WILEY ENCYCLOPEDIA OF

TELECOMMUNICATIONS

VOLUME 3

WILEY ENCYCLOPEDIA OF TELECOMMUNICATIONS

WILEY ENCYCLOPEDIA OF

TELECOMMUNICATIONS

VOLUME 3

John G. Proakis
Editor

WILEY-INTERSCIENCE

A John Wiley & Sons Publication

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

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

Library of Congress Cataloging in Publication Data:

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

2002014432

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

L

LAND-MOBILE SATELLITE COMMUNICATIONS*

JEFFREY B. SCHODORF
MIT Lincoln Laboratory
Lexington, Massachusetts

1. INTRODUCTION

Since the early 1990s there has been significant progress in the development of land-mobile satellite communications (LMSC) systems and technology. In general, LMSC service providers have struggled to compete with their terrestrial mobile wireless counterparts. However, few doubt that LMSC systems have a meaningful role to play in the quest for global wireless access in the twenty-first century. The key advantage enjoyed by satellite communications systems is their ability to cover broad geographic areas, substantially decreasing the terrestrial infrastructure required and potentially simplifying issues relating to the coordination of this infrastructure for tasks such as channel assignment and handover. Moreover, a sufficient amount of spectrum has been allocated to LMSC systems such that they represent a good choice for the delivery of broadband services such as multimedia. Of course, LMSC systems are not perfect. While fewer satellites may be required to cover an area, satellites are very expensive to build and deploy relative to terrestrial base stations. Moreover, depending on operating frequency, significant channel impairments must be overcome in LMSC systems. Nonetheless, the potential of LMSC systems ensures they will remain an area of intense research and development for the foreseeable future.

The purpose of this article is to describe in moderate detail technical issues surrounding LMSC systems. Where appropriate, references are cited so that the interested reader can pursue these topics further. In Section 2 a brief description of several existing and planned LMSC systems is given. These systems are categorized loosely by their orbital type and are further subdivided according to the services they provide. Section 3 discusses propagation issues, including path loss, signal fading, shadowing, and the effects of directional antenna mispointing. Strategies for dealing with channel impairments are discussed in Section 4. These approaches fall into one of two main categories: error control techniques such as forward error correction (FEC) coding and automatic repeat request (ARQ) protocols, and diversity combining methods. In LMSC systems, satellite resources are typically a limiting factor. Hence, efficient use of these resources is critical. Section 5 addresses this issue with a discussion of multiple access schemes. Finally, network aspects of LMSC systems are discussed in Section 6. The primary emphasis of this section is the issue of internetworking LMSC and terrestrial data networks.

* This work was sponsored by the Department of the Army under A/F Contract F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States government.

2. LAND-MOBILE SATELLITE SYSTEMS

Land-mobile satellite systems come in a variety of orbital configurations, including geostationary or geosynchronous earth orbit (GEO), medium earth orbit (MEO), and low earth orbit (LEO). GEO systems operate at an altitude of 35,786 km, and have an orbital period of 24 hs. MEO systems have altitudes ranging from 5000 to 10,000 km and have orbital periods of 4–6 hs. LEO satellites orbit at altitudes from 500 to 1500 km with periods of approximately 2 hs. Orbital mechanics will not be addressed here, but thorough treatments of this topic may be found in Refs. 1 and 2. Services provided by land mobile satellite systems include navigation, fleet management, broadcast, and duplex voice and data communications, the primary service of interest in this article.

LMSC systems are characterized by a forward path and a reverse path between two user terminals. Each path comprises an uplink between the transmitting terminal and the satellite, and a downlink between the satellite and the receiving terminal, as depicted in Fig. 1. Traditionally, LMSC systems satellites have been transponders (sometimes referred to as "bent pipes"), where the received uplink signal is simply amplified and translated to a downlink frequency. More recent systems have begun to employ processing satellites, where in addition to amplification and frequency translation, additional processing, such as demodulation, decoding, and remodulation is performed.

Table 1 summarizes the nomenclature used to categorize the various operating frequency bands of LMSC and other wireless communications systems. Original spectrum allocations for LMSC systems were in the L and

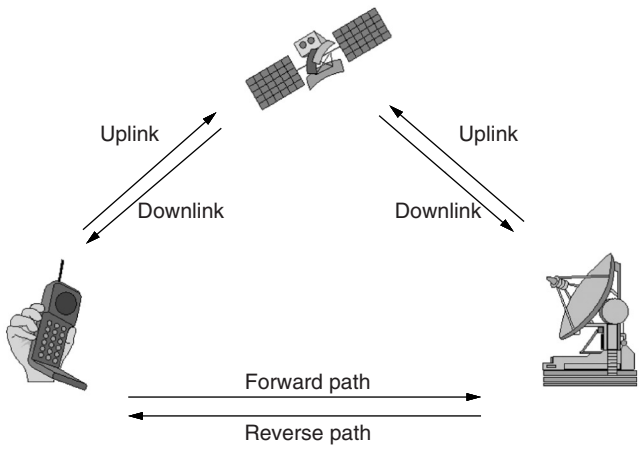


Figure 1. LMSC link nomenclature.

Table 1. Frequency Band Designations

Band	Frequency Range (MHz)
P	225–390
L	390–1550
S	1550–3900
C	3900–8500
X	8500–10,900
Ku	10,900–17,250
Ka	17,250–36,000
Q	36,000–46,000
V	46,000–56,000
W	56,000–100,000

S bands, where most systems continue to operate today. However, demand for bandwidth has resulted in additional allocations at higher frequencies. In many cases, these allocations are for fixed systems. However, at the higher frequencies terminals are typically small, and thus portable. Moreover, the distinction between the services provided by fixed and mobile systems is becoming increasingly vague. For example, consider the situation today where commercial vendors supply antenna and positioning systems that allow land mobile platforms to acquire and track DirecTV, a fixed service system. At present there is at least one operational or planned satellite system in each of the bands in Table 1, with the exception of W band.

Because of their 24-h orbital period, GEO satellites appear stationary above the equator to an observer on earth. These systems are well suited to broadcast services because a constellation size of only three or four satellites provides total earth coverage. Examples of GEO broadcast systems include DirecTV and the relatively new XM satellite radio service. For voice and data service, GEO systems have the attractive feature that no handover between satellites is necessary. On the other hand, because of their high altitude, GEO systems have long propagation delays (i.e., ~240 ms, one-way) relative to their MEO and LEO counterparts. Numerous GEO systems have been deployed for the delivery of duplex voice and data services. For example, the INMARSAT-M system provides 4.8 kbps (kilobits per second) voice capability, 2.4 kbps fax service, and 1.2–2.4 kbps data services. In addition to land-mobile terminals, INMARSAT-M also supports maritime users. Higher-data-rate systems such as the INMARSAT-4, which will support mobile communications services at rates of 144–432 kbps, are planned for the near future.

The lower orbital altitude of MEO satellites implies that they move across the sky relative to a fixed point on earth. This movement necessitates handovers between satellites. Generally, connection to the terrestrial infrastructure is achieved with a reasonable number of earth stations, or gateways. Hence, intersatellite links (ISLs) are not typically employed in MEO systems. The Intermediate Circular Orbits (ICO) system is an example of a planned MEO LMSC system. Scheduled to launch in 2004, ICO will deliver 4.8 kbps voice service, 2.4–9.6 kbps data service to handheld terminals, and 8–38.4 kbps data service to land mobile terminals. The popular Global Positioning System (GPS) is another example of a MEO satellite system, although GPS is a navigation system, as opposed to a duplex communications system.

LEO systems represent the most recent architectural system concept in LMSC systems, as well as the most complex. LEO satellites have the shortest orbital period of the three configurations discussed here. The fact that they pass rapidly over a fixed point on earth (e.g., a typical LEO satellite is in a user's field of view for 8–12 mins) implies that sophisticated handover procedures are necessary. Moreover, to connect to the terrestrial infrastructure, either a significant number of gateways are required, or else ISLs must be used. On the other hand, LEO systems offer superior delay performance and suffer less propagation loss relative to MEO and GEO systems. LEO systems are generally categorized as either "little LEO" systems or satellite personal communication networks (S-PCNs), sometimes called "big LEO" systems. Little LEO systems provide nonvoice, low-bit-rate mobile data and messaging services. Orbcomm is an example of a little LEO system currently in operation. Orbcomm offers 2.4–4.8 kbps data service to fixed and mobile users. Iridium and Globalstar are examples of S-PCNs. More information on these and other LMSC systems can be found in the literature [2,3].

3. CHANNEL CHARACTERISTICS

Fundamental to the design of any communications system is an accurate understanding of the channel over which the communications signals will be propagating. While the propagation modeling field is relatively mature for terrestrial wireless communications systems [4], it continues to be an area of active research in LMSC systems, especially at higher frequencies. In this section, numerous LMSC channel characteristics will be discussed, including random noise, path loss, weather and atmospheric effects, signal fading, shadowing, and fluctuations due to spatial tracking errors in LMSC systems that use directional antennas.

3.1. Random Noise

The dominant noise source in a communications system is usually thermal noise generated at the input to the receiver. A common practice is to lump all other noise sources together with the thermal noise and represent them as a single source added directly to the received signal. This collection of random noise is typically assumed to follow a Gaussian distribution. In digital communications systems, the effects of noise are to introduce bit errors in the received signal. The probability of error, or bit error rate (BER) in a digital communications system, is generally parameterized by the signal-to-noise ratio (SNR) per bit, or bit energy : noise ratio, E_b/N_0, where $N_0/2$ is the noise power spectral density. The relationship between BER and E_b/N_0 depends on the modulation scheme employed. Most LMSC systems use some form of constant-envelope modulation so that transmit amplifiers can be operated at saturation without introducing

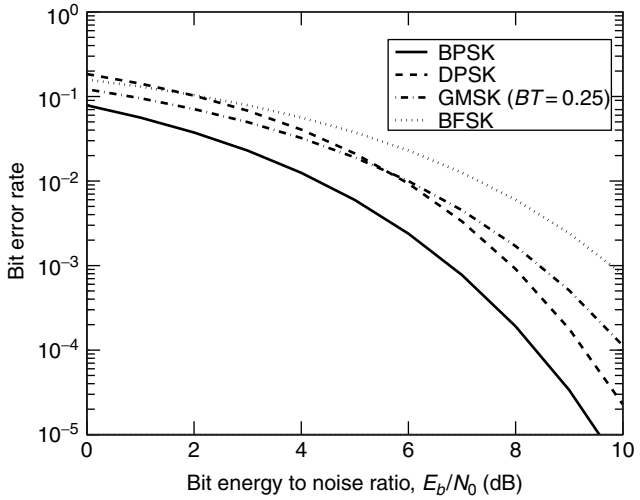


Figure 2. Performance of several modulation schemes for the Gaussian channel.

distortion into the modulated signal. Common modulation schemes include phase shift keying (PSK), frequency shift keying (FSK), and continuous-phase modulation (CPM). These techniques are described in Refs. 5 and 6, where BER expressions are derived as a function of E_b/N_0. For reference, Fig. 2 summarizes the BER performance of several modulation schemes, including binary PSK (BPSK), differential PSK (DPSK), binary FSK (BFSK), and a CPM scheme known as *Gaussian minimum shift keying* (GMSK). The corresponding equations for BER are given below:

$$\mathrm{BER}_{\mathrm{BPSK}} = Q\left(\sqrt{\frac{2E_b}{N_0}}\right) \tag{1}$$

$$\mathrm{BER}_{\mathrm{DPSK}} = \frac{1}{2}e^{-E_b/N_0} \tag{2}$$

$$\mathrm{BER}_{\mathrm{BFSK}} = Q\left(\sqrt{\frac{E_b}{N_0}}\right) \tag{3}$$

$$\mathrm{BER}_{\mathrm{GMSK}} \approx Q\left(\sqrt{\frac{2\alpha E_b}{N_0}}\right) \tag{4}$$

where $Q(x) = \int_x^\infty 1/\sqrt{2\pi}e^{-u^2}\,du$ is the Gaussian Q function. The term α in the approximation for BER in GMSK systems is a scalar that depends on the time–bandwidth product, BT. For $BT = 0.25$, $\alpha = 0.68$ [6].

In order to asses the quality of a satellite link between two terminals, both uplink and downlink must be considered. In transponded satellite systems, the following relationship holds between the bit energy to noise of the total link, $(E_b/N_0)_{\mathrm{tot}}$, and the bit energy to noise ratio of the uplink and downlink, assuming that the transponder and receiving terminal bandwidths are the same:

$$\left(\frac{E_b}{N_0}\right)_{\mathrm{tot}} = \frac{(E_b/N_0)_{\mathrm{ul}}(E_b/N_0)_{\mathrm{dl}}}{(E_b/N_0)_{\mathrm{ul}} + (E_b/N_0)_{\mathrm{dl}}} \tag{5}$$

where $(E_b/N_0)_{\mathrm{ul}}$ and $(E_b/N_0)_{\mathrm{dl}}$ represent the bit energy : noise ratio of the uplink and downlink, respectively. The BER of the total link is then based on $(E_b/N_0)_{\mathrm{tot}}$ and the modulation scheme used. In processing satellites, the uplink and downlink can be analyzed separately, and the following approximation holds:

$$(\mathrm{BER})_{\mathrm{tot}} \approx (\mathrm{BER})_{\mathrm{ul}} + (\mathrm{BER})_{\mathrm{dl}} \tag{6}$$

where $(\mathrm{BER})_{\mathrm{tot}}$ is the BER of the total link, $(\mathrm{BER})_{\mathrm{ul}}$ is the uplink BER, and $(\mathrm{BER})_{\mathrm{dl}}$ is the downlink BER.

3.2. Path Loss

In LMSC systems, path loss arises from several sources. Free-space path loss arises in wireless communications systems due to the spatial dispersion of the radiated power, and is quantified as follows:

$$L_0 = 10\log\left(\frac{4\pi d}{\lambda}\right)^2 \tag{7}$$

where L_0 is the free-space path loss in dB, d is the distance between the transmitter and receiver, and λ is the signal wavelength. Figure 3 illustrates free-space loss for several frequencies as a function of satellite altitude, d_a. Note, however, that in LMSC systems the distance between a user terminal and a satellite is not the same as the satellite altitude. The user's location on the earth must be considered as well. A conceptually simple way to consider the problem is to treat the total distance between a satellite and user terminal as the sum of two terms:

$$d_{sr} = d_a + d_e \tag{8}$$

where d_e is an elevation dependent term that, when added to the satellite altitude d_a, yields the total distance between the satellite and user terminal, referred to as the *slant range*. The additional free space loss, in decibels, due to the slant range may be expressed as

$$\Delta L_{sr} = 20\log\frac{d_{sr}}{d_a} \tag{9}$$

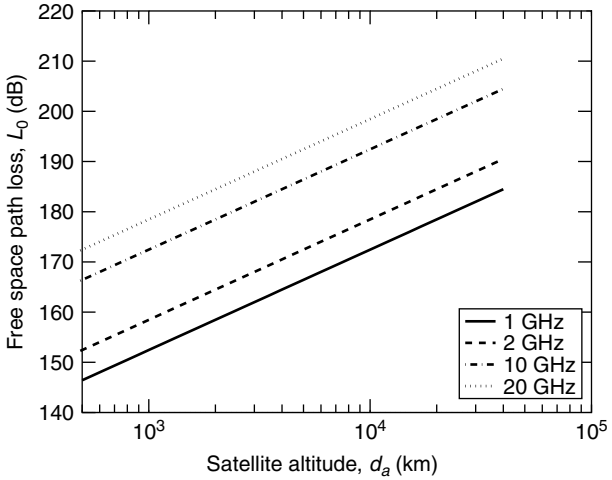


Figure 3. Free-space path loss as a function of satellite altitude d_a for several different operating frequencies.

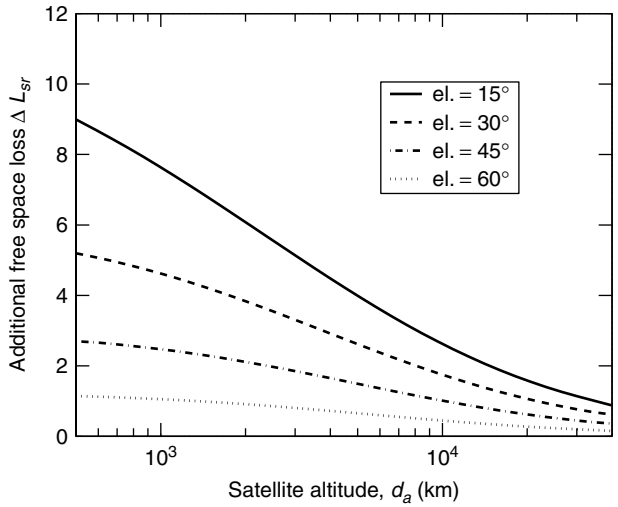


Figure 4. Additional free-space path loss due to slant range.

Figure 4 gives ΔL_{sr} as a function of satellite altitude for several different elevation angles.

Weather and atmospheric effects are another factor that contribute to path loss in LMSC systems. Above X band, signal scattering and absorption by water droplets are the dominating factors. In general, attenuation due to rain is a function of the amount of rainfall, drop size, temperature, operating frequency, and pathlength through the rain. Results from measurement campaigns at millimeter wavelengths [7,8] suggest that "typical" rain rates of 20 mm/h yield losses on the order of 2 dB/km. Obviously, rain rate statistics will vary from region to region, thus affecting average losses.

3.3. Multipath Fading

The term *multipath fading* is used to describe the phenomenon whereby multiple, reflected versions of a transmitted signal (i.e., multipath components) combine at a receiver in either a constructive or destructive fashion depending on their relative amplitudes and phases. When either the transmitter or receiver is in motion, the dynamic, random combining of multipath components leads to received signals that can vary by several tens of decibels with relatively small changes in spatial location. Statistically, multipath fading may be treated as a random process. In the event that no single dominant propagation path exists, the fluctuations in received signal power S are described by the central chi-square distribution [5]:

$$p(S \mid S_0) = \frac{1}{S_0} \exp\left(-\frac{S}{S_0}\right) \tag{10}$$

where S_0 is the mean received signal power, due entirely to multipath. This class of channel is often referred to as a *Rayleigh channel* because the received signal envelope follows the Rayleigh distribution. This statistical model holds well in practice for terrestrial microwave cellular systems, where typically no line-of-sight (LoS) path between transmitter and receiver exists [9]. In LMSC systems, a LoS path often exists in addition to the multipath. In this case, the received signal power, S, is described by the noncentral chi-square distribution [5]

$$p(S \mid A, \sigma_d^2) = \frac{1}{\sigma_d^2} \exp\left\{-\frac{A^2 + 2S}{2\sigma_d^2}\right\} I_0\left(\sqrt{2S}\frac{A}{\sigma_d^2}\right) \tag{11}$$

where A is the amplitude of the LoS component, σ_d^2 is the diffuse signal power, and I_0 is the modified zeroth-order Bessel function of the first kind. This class of channel is often referred to as a Ricean channel because the received signal envelope follows a Ricean distribution. Ricean channels are frequently parameterized by the *Rice factor:*

$$c = \frac{A^2}{2\sigma_d^2} \tag{12}$$

The Rice factor is simply the ratio of the power in the direct and multipath components. When no LoS component exists (i.e., $A = 0$), $c = 0$ and (11) reduces to (10) with the mean received signal power given by $S_0 = \sigma_d^2$. When $c = \infty$, the channel does not exhibit fading. Figure 5 compares

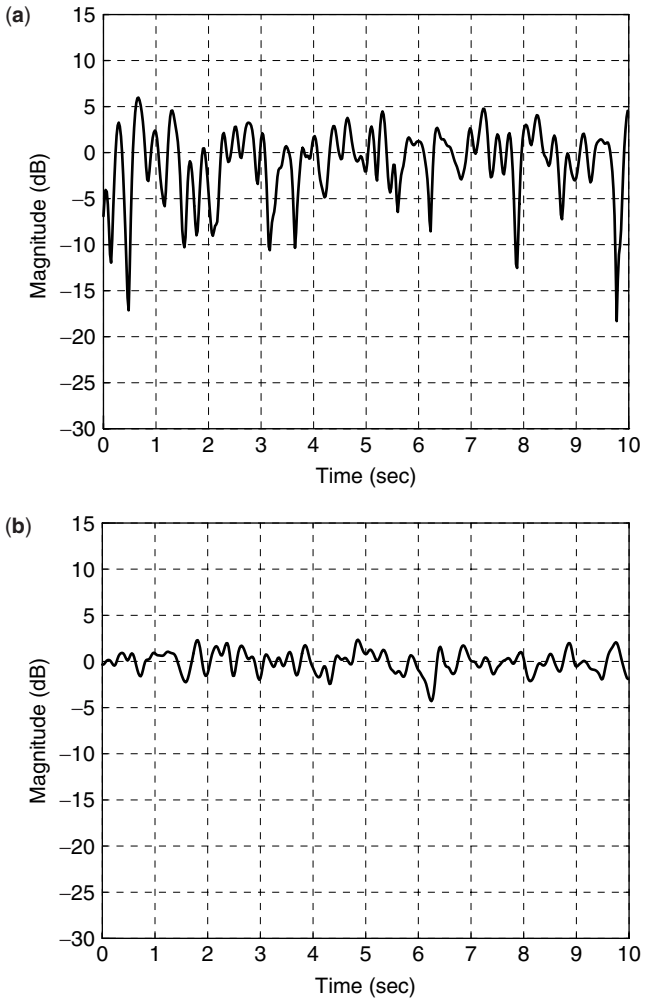


Figure 5. Ricean fading envelopes for (**a**) $c = 10$ dB and (**b**) $c = 20$ dB.

fading signal envelopes for two values of the Rice factor: $c = 10$ dB, and $c = 20$ dB.

In LMSC systems, the Rice factor depends on a number of parameters, including operating frequency, elevation angle, and antenna type. In general, systems that operate at higher frequencies will experience less multipath due to their use of directive antennas and the tendency of shorter wavelengths to scatter off objects in the propagation path. Hence, these systems are typically characterized by larger Rice factors. For example, whereas Rice factors reported for the L-band system studied by Lutz et al. [10] average approximately 10 dB, the Rice factors reported from NASA's Advanced Communications Technology Satellite (ACTS) propagation experiments [11–13], conducted at 20 GHz, average more than 20 dB. Vogel and Goldhirsh [14] examined the multipath fading phenomenon in detail at low elevation angles in unshadowed LOS environments for the INMARSAT LMSC system, which operates at L band. Experimental results show that at elevation angles from 7° to 14°, fades exceeding 7 dB occur for approximately 1% of the driving distance. Moreover, the authors note that the fading is typically dominated by a single multipath reflection from a nearby terrain feature.

The effects of multipath fading on the BER of an LMSC system are quite severe. Because of the variations in received signal strength, the average bit energy : noise ratio $\overline{E_b}/N_0$ must be used in characterizing the average BER performance. Proakis has derived [5] expressions for average BER as a function of $\overline{E_b}/N_0$ for BPSK and DPSK modulations in a Rayleigh fading environment:

$$\overline{\mathrm{BER}_{\mathrm{BPSK}}} = \frac{1}{2}\left(1 - \sqrt{\frac{\overline{E_b}/N_0}{1+\overline{E_b}/N_0}}\right) \tag{13}$$

$$\overline{\mathrm{BER}_{\mathrm{DPSK}}} = \frac{1}{2\left(1+\overline{E_b}/N_0\right)} \tag{14}$$

where $\overline{\mathrm{BER}}$ denotes average BER. Average BER performance in Ricean fading environments was examined in a 1995 article [15]. Figure 6 summarizes these results with BPSK and a couple of different Rice factors. The average BER of BPSK for the Gaussian and Rayleigh channels are also included for reference. Note the significant difference in required $\overline{E_b}/N_0$ necessary to achieve a given BER in the presence of multipath fading. For example, more than 10 dB separates the Gaussian channel and the Ricean channel with $c = 7$ dB at $\overline{\mathrm{BER}} = 1e-4$.

3.4. Shadowing

Signal shadowing is caused when relatively large-scale objects, such as buildings or terrain features, either partially or completely intersect the propagation path. Although difficult to model mathematically, variations in the received signal power S_0, caused by shadowing have been observed to follow a lognormal distribution (i.e., a distribution whose values, when plotted on a log scale, appear Gaussian)

$$p(S_0) = \frac{10}{\sqrt{2\pi}\sigma \ln 10}\frac{1}{S_0}\exp\left[-\frac{(10\log S_0 - \mu)^2}{2\sigma^2}\right] \tag{15}$$

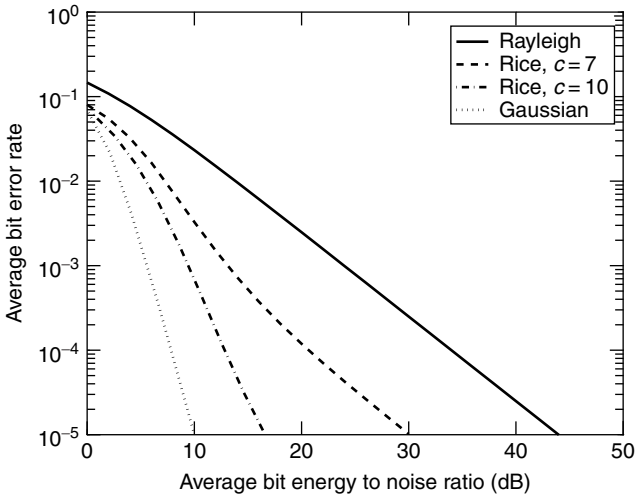


Figure 6. Performance of BPSK in the presence of Ricean and Rayleigh multipath fading.

with a mean μ and standard deviation σ that depend on the carrier frequency and environment [4]. Above X band, the effects of shadowing tend to be more severe, due to the absence of received multipath energy. In these systems the LoS path is critical, and obstruction of this path results in a nearly total loss of received signal power (i.e., signal blockage). Published results from the ACTS mobile propagation experiments [11–13] at 20 GHz report typical means from −15 to −20 dB and standard deviations in the range of 5–10 dB. It is also important to note that path losses due to foliage are significant enough to classify trees as objects that give rise to shadowing in systems that operate above X band. According to certain foliage path loss models [4,16], 5–10 ms of foliage is sufficient to yield losses on the order of 10–15 dB at 20 GHz.

Lutz et al. [10] proposed a total shadowing model (TSM) that effectively combines the densities given by (10), (11), and (15) to describe a LMSC propagation channel at L band. A timeshare parameter $0 \le X \le 1$ was introduced such that a fraction X of the time the received signal power is unaffected by shadowing and described by (11), while the remaining fraction, $(1-X)$, of the time the LoS component is totally blocked (i.e., $A = 0$) and the received signal power follows (10) with the average power S_0, given by the lognormal density in (15). Expressed mathematically. The TSM is given by

$$\begin{aligned} p(S) &= X\,p(S \mid A, \sigma_d^2) + (1-X)\int_0^\infty p(S \mid S_0)\,p(S_0)\,dS_0 \\ &= X\frac{1}{\sigma_d^2}\exp\left\{-\frac{A^2+2S}{2\sigma_d^2}\right\} I_0\left(\sqrt{2S}\frac{A}{\sigma_d^2}\right) \\ &\quad + (1-X)\int_0^\infty \frac{1}{S_0}\exp\left(-\frac{S}{S_0}\right) \\ &\quad \times \frac{10}{\sqrt{2\pi}\sigma \ln 10}\frac{1}{S_0}\exp\left[-\frac{(10\log S_0 - \mu)^2}{2\sigma^2}\right] dS_0 \end{aligned} \tag{16}$$

According to this equation, the fading behavior of the channel consists of two dominant modes or states. In

the unshadowed state (i.e., the "good" channel state) the channel is characterized by the presence of a LoS component, which implies high received power and Ricean fading, while in the shadowed state (i.e., the "bad" channel state) the channel is characterized by the absence of a LoS component, which implies low received power and Rayleigh fading. The timeshare parameter X is a long-term average that describes the fractional amount of time spent in each state. The short-term characteristics of the switching process are accurately described by a two-state Markov model [10]. The situation is depicted in Fig. 7. When the channel is in the good state G, there is a probability p_{GG} associated with remaining in that state and a crossover probability p_{GB} associated with the transition to the bad state B such that $p_{GG} + p_{GB} = 1$. Likewise, there is a probability p_{BB} associated with remaining in the bad state and a probability p_{BG} associated with switching from the bad state to the good state such that $p_{BB} + p_{BG} = 1$. According to the model, the mean duration, in bits, of a good or bad channel state is given by

$$G_b = \frac{1}{p_{GB}}$$
$$B_b = \frac{1}{p_{BG}} \tag{17}$$

and the probability that a good or bad channel state lasts longer than n bits is given by

$$p_G(> n) = p_{GG}^n$$
$$p_B(> n) = p_{BB}^n \tag{18}$$

In addition, the timeshare parameter X can be expressed in terms of the Markov model parameters:

$$X = \frac{G_b}{G_b + B_b} = \frac{p_{BG}}{p_{BG} + p_{GB}} \tag{19}$$

In [10] the Markov model parameters were estimated by fitting the statistics to actual recorded data. The validity of the model described by (16)–(19) for the Ka-band LMSC channel was verified over the course of the ACTS mobile propagation experiments and values for the various model parameters, including X, c, μ, σ, p_{GG}, p_{BB}, G_b, and B_b were reported by Rice [11].

In addition to the TSM, numerous other statistical models have been proposed to describe the LMSC channel. Loo [17] presented a statistical model for L-band LMSC systems. Expressions for level crossing rate and average fade duration are derived and compared to measured data, where reasonably good agreement is observed. Another L-band statistical model, proposed for nongeostationary (i.e., LEO and MEO) LMSC systems has been presented [18]. The model is tunable over a range of environments. Moreover, comparisons to real data are used to derive empirical formulas for the model parameters for several different elevation angles. Finally, a comprehensive statistical model has been proposed [19]. This model is intended to cover a broad range of operating environments and frequencies. In addition, the model can be used to generate time series for LMSC signal features that include amplitude, phase, instantaneous power delay profiles, and Doppler spectra.

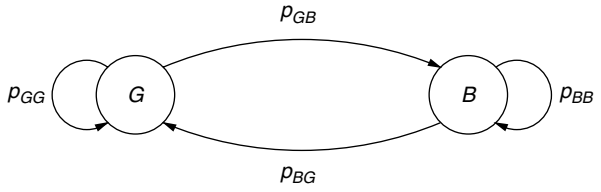


Figure 7. Two-state Markov model that describes the switching process in a LMSC channel.

3.5. Fading Due to Antenna Mispointing

In many LMSC systems directional antennas are typically employed for their high gain. One challenge associated with this practice is maintaining accurate pointing of the receive terminal's directive antenna despite the vehicle dynamics. Regardless of the pointing system used, there will always be residual mispointing error. Characterizing the fluctuations in received signal strength due to antenna mispointing is difficult because of the wide variety of factors that contribute to pointing errors. These factors include terrain, vehicle type and speed, antenna beamwidth, and the antenna controller.

In the ACTS system, mobile propagation experiments were conducted at 20 GHz using the ACTS mobile terminal (AMT). The AMT uses a mechanically steered elliptically shaped reflector antenna with dimensions of approximately 6×2.5 ins. More details on the AMT antenna and tracking system can be found papers by Densmore and others [20,21]. Rice et al. [22] experimentally characterized mispointing error for the AMT. Specifically, measurements of the vehicle pitch, roll, and heading were taken at 0.1-mi intervals along a specific route traveled by the terminal. The error associated with these measurements was used to upper-bound the azimuth and elevation angle mispointing errors at 3.9° and 3.3°, respectively. Finally, through logarithmic interpolation of the antenna gain pattern data, the loss in received signal power due to antenna mispointing was calculated to be on the order of 1.5 dB. These results were then used to rationalize the 1–2-dB variations in received signal power observed during previous AMT runs. Rice and Humphreys [12,13] used data from a wider range of experiments to develop a somewhat ad hoc statistical characterization of antenna mispointing with the AMT. Specifically, a bimodal density function was proposed since this model was observed to fit the experimental data.

A probabilistic analysis of antenna mispointing was presented in an earlier work in which, Gaussian distributed mispointing errors in the azimuth and elevation directions were assumed [23]. This assumption reflects heuristic observations made with pointing systems designed to support Ka-band LMSC systems over rugged terrain. The analysis is of sufficient generality to apply to a range of pointing systems and antenna types since the case where unequal azimuth and elevation mispointing variance are as well as the equal-variance

case examined. Probability density functions (PDFs) for the total mispointing error (i.e., the vector sum of the azimuth and elevation mispointing errors) are given. In addition, the antenna mispointing PDF is used to generate the PDF for received signal loss, from which the average BER is easily computed.

4. ERROR MITIGATION

Consider the following, somewhat simplified, link budget equation between a ground terminal and satellite:

$$\frac{E_b}{N_0} = \text{EIRP} - L + \frac{G}{T} - k_b - R - M \text{ dB} \qquad (20)$$

where $\text{EIRP} = P_t + G_t$ is the effective isotropic radiated power of the transmitter (i.e., the sum, in decibels, of the transmit power and antenna gain) and L is a loss term that accounts for free space and other path losses. The term G/T, where G is the receive antenna gain and T is the effective noise temperature, including the noise temperature of the antenna and the effective noise temperature of the receiver low-noise amplifier (LNA), is often referred to as the *figure of merit* of the receiver. The constant k_b is Boltzmann's constant, $k_b = -228.6$ dBW K^{-1} Hz^{-1} (i.e., decibel watts per degree Kelvin per hertz). The transmission rate, in bits per second (bps), that can be supported by the channel, is given by R. Finally, the link margin, M, is a contingency included to overcome implementation losses and channel effects such as weather, multipath fading, shadowing, and antenna mispointing.

Table 2 illustrates link budget parameters for a typical processing LEO satellite uplink, such as Iridium. In this example, $E_b/N_0 = 8$ dB. Assuming BPSK modulation, Fig. 6 shows that a BER of 2×10^{-4} is achieved for this scenario with full margin. However, in order to achieve the same BER in the presence of Ricean fading where $c = 7$ dB, an additional 10 dB of signal power is required, leaving only 2 dB of link margin available to address other losses. Clearly, any scheme that is capable of reducing the required channel bit energy : noise ratio, $(E_b/N_0)_{\text{req}}$ for a given performance (i.e., BER) level, is of interest. After all, if $(E_b/N_0)_{\text{req}}$ can be reduced, the savings can be applied directly to reduce the transmitter power, antenna gain, or other parameters, or to increase the channel data rate. In this section, two approaches to reducing $(E_b/N_0)_{\text{req}}$ are discussed: error control in the form of FEC coding and ARQ protocols, and diversity combining.

Table 2. LEO Satellite Uplink Budget

Terminal transmit power	$P_t = 1$ W $\equiv 0$ dBW
Terminal antenna gain	$G_t = 2$ dBi
Terminal EIRP	2 dBW
Free-space path loss (2 GHz operating frequency, 800 km orbit altitude)	$L_0 = 166.1$ dB
Additional loss due to 30° satellite elevation (i.e., slant range)	$\Delta L_{sr} = 5.1$ dB
Total budgeted path loss	$L = 171.2$ dB
Gain of satellite receive antenna	26 dBi (edge of coverage)
Antenna noise temperature	290 K
LNA noise temperature	75 K
Effective noise temperature	$T = 365$ K $\equiv 25.6$ dBK
Satellite G/T	$G/T = 0.4$ dB
Boltzmann's constant	$k_b = -228.6$ dBW/K/Hz
Channel rate	9.6 Kbps $\equiv 39.8$ dBHz
Link margin	12 dB

4.1. FEC Coding and ARQ Protocols

The main idea behind power-efficient FEC coding is that the introduction of redundancy into the transmitted bit stream can be exploited by a receiver to correct bit errors caused by channel impairments. The redundancy comes in the form of additional (i.e., parity) bits that are carefully selected by some encoding algorithm and inserted into the transmitted bit stream. With knowledge of the encoding algorithm, the receiver is able to reverse the encoding process, or decode the received sequence. Depending on the sophistication of the encoding/decoding procedures, errors caused by channel noise, fading, or other anomalies can be detected and/or corrected. Note also that as the name implies, FEC coding algorithms operate on the forward link only. No feedback or return link is necessary.

Numerous FEC coding strategies have been developed over the years, including block coding techniques and convolutional codes. Block coding strategies operate on fixed-size blocks of information bits. For each k-bit input block of information bits, the encoding algorithm produces a unique n-symbol output block (i.e., codeword). Note that for a given channel rate, in bps, the *information rate*, that is, the channel capacity devoted to carrying the information bits, is decreased by a factor $R_c = k/n$, called the *code rate*. As with block codes, convolutional codes can be used to generate n-symbol outputs for k-bit inputs yielding a rate $R_c = k/n$ code. However, the primary characteristic that distinguishes convolutional codes from block codes is the fact that convolutional codes have memory. In other words, an n-symbol output block depends not only on the corresponding k-bit input block, but also the m previous input blocks, where m is the memory order of the encoder. Detailed treatments of these approaches are available from a variety of sources [e.g., 5,24]. In recent years, a new and extremely powerful approach to FEC coding, referred to as *Turbo coding* [25], has been introduced. At the heart of Turbo coding schemes are the concepts of code concatenation, soft-decision decoding, and iterative decoding [26]. In general, Turbo coding schemes are more computationally demanding than either block or convolutional codes. Within the context of LMSC systems, convolutional codes remain a popular choice because of their good performance and relatively simple implementation. However, the superior performance advantages of Turbo codes, coupled with advances in decreased-complexity implementations and ever-increasing microprocessor speeds, suggest that Turbo

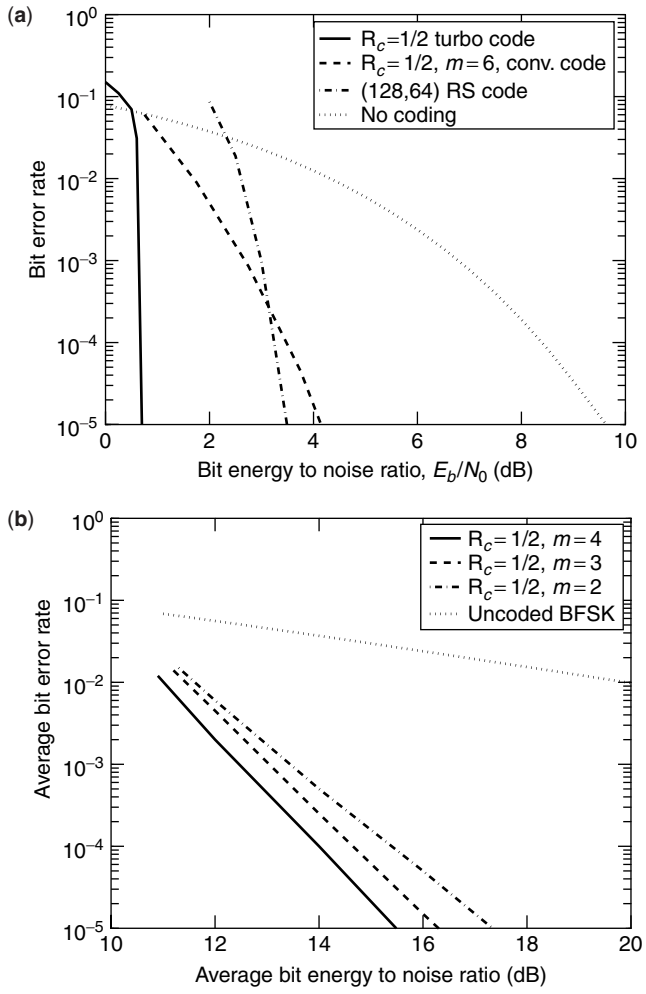


Figure 8. FEC coding performance: (**a**) a $R_c = \frac{1}{2}$ Turbo code, a $R_c = \frac{1}{2}$, $m = 6$ convolutional code (with soft-decision decoding), and a $n = 128$, $k = 64$, (i.e., $R_c = \frac{1}{2}$) Reed–Solomon (RS) block code with BPSK modulation for the Gaussian channel; (**b**) FEC convolutional coding with BFSK modulation for the Rayleigh fading channel; $R_c = \frac{1}{2}$ codes with $m = 2, 3, 4$ are compared.

codes are a logical choice for future systems. Figure 8a illustrates the performance of several different $R_c = \frac{1}{2}$ codes, including the parallel concatenated Turbo code in [25], with BPSK modulation and a Gaussian channel. Proakis [5], examined the performance of convolutional codes with noncoherent FSK modulation for Rayleigh fading channels and derived upper bounds. Figure 8b summarizes these results.

As opposed to FEC coding, ARQ schemes operate by requesting a retransmission of the codeword rather than attempting to correct it at the receiver. Of course, the existence of a feedback path (i.e., a return link) is required with such an approach. Typically, coding is still used in ARQ strategies, but only to alert the receiver to the presence of errors, not to correct them. Since the probability of an undetected error is usually much smaller than that of a decoding error, ARQ schemes are an inherently more reliable form of error control than FEC coding. Hence, these schemes are most often associated with data communications where very low error rates are required. As established previously, FEC coding is appropriate for addressing bit errors introduced by random noise and multipath fading in LMSC systems. On the other hand, ARQ protocols are well suited to situations where long deep fades, such as those caused by shadowing and signal blockage in high-frequency systems, are expected [27]. A thorough description of the main forms of ARQ, including stop and wait, go-back N, and selective repeat is available [24]. It is also possible to combine FEC coding and ARQ schemes into hybrid ARQ (HARQ) protocols that offer performance advantages as well as other desirable attributes such as rate adaptation.

4.2. Diversity Techniques

The basic idea of diversity signaling is that the effects of signal fading can be reduced by supplying the receiver with replicas of the same transmitted signal information over independently faded channels. Through proper selection or combining of these replicas, the likelihood that the receiver experiences a deep fade is reduced considerably. For example, if the probability that any one signal fades below some threshold is p_f, then the probability that D independently faded replicas of the same signal will simultaneously fade below this threshold is given by p_f^D. The optimum strategy for diversity reception is that the D signal replicas be combined coherently in proportion to their received SNR. This type of combining scheme is often referred to as *maximal ratio combining*. A very simple approximation to the average BER of a BPSK signal in a Rayleigh fading environment with maximal ratio combining of D independently fading diversity branches is given by

$$\overline{\mathrm{BER}_{\mathrm{BPSK}}^{mr}} = \left(\frac{1}{4(\overline{E_b}/N_0)_d}\right)^D \binom{2D-1}{D} \tag{21}$$

where $\overline{\mathrm{BER}^{mr}}$ is the average BER for maximal ratio combining and $(\overline{E_b}/N_0)_d$ is the average bit energy : noise ratio received from the D diversity branches:

$$\left(\frac{\overline{E_b}}{N_0}\right)_d = \sum_{i=1}^{D} \left(\frac{\overline{E_b}}{N_0}\right)_i \tag{22}$$

and $\binom{a}{b}$ represents the number of possible combinations of a objects taken b at a time. Figure 9 summarizes the results for several values of D.

There are many ways in which diversity can be introduced into a communications system, most of which have been explored thoroughly within the context of terrestrial cellular systems [9]. These techniques include time diversity, frequency diversity, polarization diversity, and spatial diversity. With spatial diversity, multiple receive antennas are spaced far enough apart so as to yield sufficiently low correlation between their outputs. In environments where significant multipath energy is received from numerous different directions, spacings on the order of $\lambda/4$ are appropriate [9]. In situations where

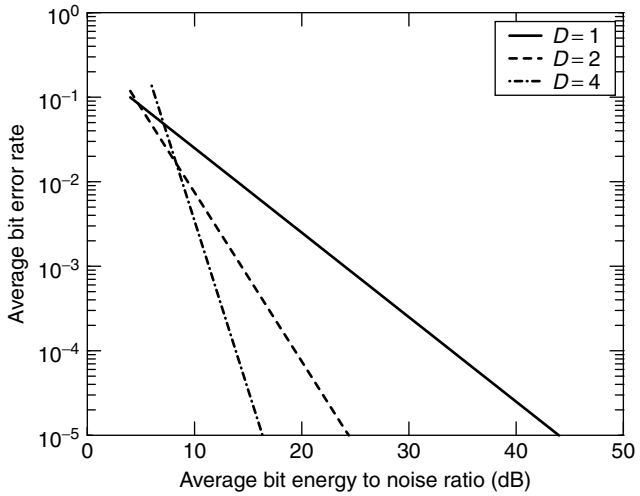


Figure 9. Performance of D-branch maximal ratio diversity combining with BPSK signaling in a Rayleigh multipath fading channel.

received power is confined to a relatively narrow sector in space, such as those where a strong LoS component exists, wider spacings are required. Spatial diversity is especially popular in cellular radio because no modifications to the waveform or transmitter are required. Instead, only $D - 1$ additional antennas, including RF chains, and combining logic are required at the receiver.

An idea similar to spatial diversity has received attention in the LMSC community. The concept is referred to as *satellite diversity*, and is applicable in situations where multiple satellites are potentially within the field of view of a terminal (e.g., LEO and MEO systems). With satellite diversity, terminals communicate with the satellite that provides the best link quality [28], thus improving link availability in the presence of signal shadowing and blockage. In some cases, multiple satellites may transmit the same signal to a user terminal, where these downlinks are combined for improved performance. Obviously, the effectiveness of satellite diversity depends on a variety of factors, including the satellite constellation and the propagation environment. Results from a measurement campaign in Japan [29] suggest that fade margins can be reduced by approximately 10 dB in urban environments using twofold satellite diversity in the Globalstar (i.e., LEO) satellite system. Satellite diversity is an integral part of both the Globalstar and ICO system concepts. The XM satellite radiobroadcast system uses a GEO constellation but achieves the same effect as satellite diversity through the use of terrestrial repeater stations located in heavily shadowed environments such as dense urban areas.

5. MULTIPLE ACCESS

Multiple access is concerned with ways in which a group of users share a common communications resource in an efficient manner. With respect to LMSC systems, the common communications resource is the satellite. Because the satellite resource is limited, and because upgrading this resource is difficult once the satellite is in orbit, efficient multiple-access schemes represent a critical component to LMSC systems. Satellite multiple-access schemes fall into one of four categories [30]: fixed-assignment techniques, random-access methods, demand assignment protocols, and adaptive assignment protocols.

Fixed assignment schemes are most appropriate in situations where users' needs are such that resources should be dedicated for a relatively long period of time (at least long enough to justify the overhead associated with allocating and deallocating the resources), such as voice circuits in a satellite telephony system. Frequency division multiple access (FDMA), time-division multiple access (TDMA), and code-division multiple access (CDMA) are the basic forms of fixed assignment. With these schemes, the satellite resource is partitioned into orthogonal, or quasiorthogonal segments, referred to as *channels*. In FDMA, the channels are fixed slices of bandwidth, to which users are granted exclusive use for the duration of their call, or session. In TDMA, channels are created through the use of a framing structure that divides the resource into time slots. Time slots are then assigned to users whereby they are allowed access to the entire bandwidth for the duration of their slot. Time slots are typically quite short but repeat on a regular basis so that from the users' perspective a constant data rate is achieved. Finally, with CDMA, channels are associated with special periodic spreading codes, which are used to modulate the users' bit streams, resulting in bandwidth expansion. Hence, with CDMA, users overlap one another in both frequency and time. However, provided the codes are orthogonal, they are distinguished at the receiver by correlating with the desired user's code. Figure 10 illustrates the three concepts. FDMA and TDMA have been used in LMSC systems for quite some time. However, CDMA is gaining popularity [31] and is currently used in the Globalstar system.

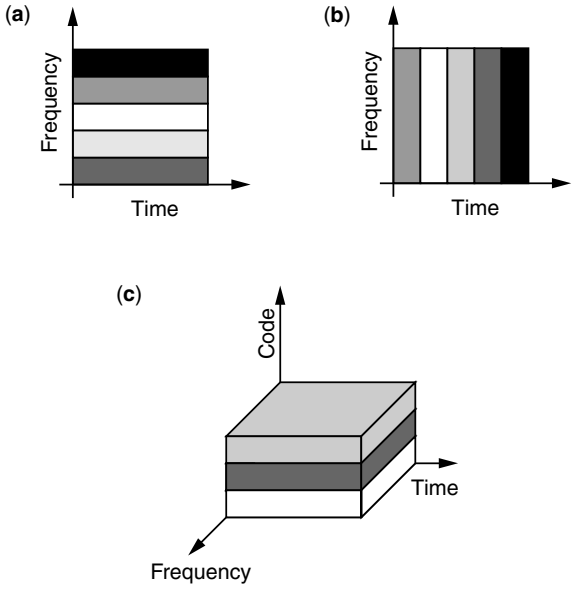


Figure 10. Fixed assignment multiple access schemes: (**a**) FDMA; (**b**) TDMA; (**c**) CDMA.

Random access schemes are appropriate in situations where user traffic is packetized and bursty. With random access, users contend for the common communications resource. A user who has a packet to send simply sends it. The user then monitors the satellite downlink to determine whether the packet was correctly received, and hence forwarded, by the satellite. Or, an acknowledgment scheme may be used whereby the receiving terminal notifies the sender via an acknowledgment message (ACK). If the packet is not heard on the downlink, or the ACK is not received, a collision with another user's packet is assumed and the packet is resent after a random delay. ALOHA and slotted ALOHA are examples of random-access protocols that operate in this manner [32]. The throughput, defined as the expected number of successful transmissions per unit time of ALOHA, is approximately 18.4%. The slotted time structure imposed by slotted ALOHA reduces the likelihood of a collision and effectively doubles the throughput to 36.8%. The obvious drawback to random-access schemes is that as traffic loading increases, so does the probability of collision, which negatively impacts performance.

Demand assignment protocols typically consist of two phases: a *reservation phase*, where resources are requested by users according to their needs; and a *communications phase*, where these resources are actually used. Reservations are typically carried out on a separate channel via random access, with the logic that reservation packets are typically short and less frequent than data messages. Packet reservation multiple access [33] is an example of a demand assignment protocol. These schemes are most appropriate when users have, on average, moderate communications requirements (e.g., occasional large file transfers). Of course, the penalty associated with demand assignment schemes is the overhead and latency (an extra round-trip delay through the LMSC system) associated with the reservation request.

Adaptive assignment protocols use a variety of means to dynamically adjust to the traffic type and load. In most cases, these schemes represent a hybrid, or superset, of other multiple access schemes. For example, the approach in [34], referred to as the "urn" scheme because of the analogy of drawing balls from an urn used in its development, adapts smoothly from the slotted ALOHA random access scheme in lightly loaded conditions to the TDMA fixed assignment scheme in heavily loaded conditions. In priority oriented demand assignment (PODA) [35], a framing structure is imposed whereby frames are subdivided into reservation and information subframes. Reservation subframes are used to make explicit reservations as in a typical demand assignment scheme. However, the protocol also supports implicit reservations whereby reservations for subsequent packets may be piggybacked onto transmitted information packets, thus improving performance for streaming traffic. In general, adaptive assignment approaches will work best when users' communications requirements are mixed, since the schemes will adapt as necessary. However, adaptive assignment protocols are also more complex relative to other multiple-access schemes.

6. NETWORK ASPECTS

Networking in LMSC systems is a relatively broad topic that has at least two major components: issues relating to LMSC cellular voice networks and those related to the internetworking of LMSC and terrestrial data networks. Only a few points will be made in regard to LMSC voice networks since there is a great deal of similarity between these systems and terrestrial cellular voice networks. The discussion on internetworking focuses exclusively on the problem of improving the performance of the standard TCP/IP protocol suite in data networks that contain both terrestrial and LMSC links.

In most respects, network procedures in LMSC cellular voice systems are quite similar to those in terrestrial cellular networks. Of course, these similarities are largely by design to promote interoperability among terrestrial and LMSC networks. In cases where distinctions occur, they are due mostly to differences in the topology of LMSC systems compared to terrestrial cellular systems, and/or the fact that satellites in LMSC systems might move relative to a fixed point on earth. For example, in terrestrial cellular systems, mobile users communicate directly with a base station that is responsible for serving a particular cell. This affiliation facilitates a number of network procedures, including those associated with mobility management (e.g., location registration and handover of live calls from one base station to another), and those associated with resource management (e.g., call setup and teardown, channel allocation, and paging). Because in terrestrial systems there is generally a one to one correspondence between a base station and a cell, the affiliation process is relatively straightforward. As users roam from cell to cell, they associate with the appropriate base station as determined by the quality of a received broadcast transmission that emanates from the various base stations in a service region. In LMSC systems, fixed earth stations, referred to as "gateways," act as the interface to the terrestrial wired infrastructure, and are similar to base stations in this regard. However, one important distinction between base stations and gateways is that whereas in terrestrial cellular systems users communicate with the base station directly, in LMSC systems users reach a gateway by communicating through a satellite link (see Fig. 1). This difference in network topology complicates the relationship between a gateway and the users it serves in a particular area, especially in LEO and MEO systems, where the satellites move relative to the gateways. For example, consider a LEO or MEO system that does not employ ISLs. In this case, the service area of a gateway, defined as the area in which both the mobile user and gateway have a simultaneous view of the same satellite, actually changes over time as the satellites move relative to the earth. In systems where ISLs are used, the service area of a gateway is completely arbitrary. In fact, it is theoretically possible that a single gateway could be used in these systems to serve all users [2].

Just as interoperability among terrestrial and LMSC voice networks is desirable, the same is true with data networks. Internetworking of disparate terrestrial

data networks is often achieved in part through the use of the TCP/IP protocol suite [32], where IP is responsible primarily for message routing and TCP includes mechanisms for flow control and error recovery. Unfortunately, the use of TCP in wireless and satellite networks generally results in suboptimal performance. The primary reason for poor TCP performance in these situations is that TCP flow control and error recovery algorithms were originally designed for use in wired systems where BERs are relatively low, latencies are short and forward and return paths are generally symmetric with respect to data rate. On the other hand, networks that include satellite links will exhibit higher BERs, longer latencies, and possibly asymmetric data rates on the forward and return paths. These differences in link conditions adversely affect the performance of TCP and result in suboptimal performance [36]. For example, because it was designed for use in terrestrial wired networks where the BERs are usually low, the TCP flow control algorithm attributes packet loss to congestion and reduces link utilization to accommodate the situation. In LMSC systems, where packets are often lost because of bit errors as opposed to congestion, the resulting reduction in utilization is unwarranted and represents an inefficiency. The relatively long round-trip time in LMSC systems adversely impacts both TCP flow control and error recovery, especially when high data rates are used (i.e., the system has a large bandwidth–delay product). With respect to flow control, TCP uses an algorithm known as "slow start," whereby the packet transmission rate is increased gradually on the basis of received acknowledgments. In large bandwidth–delay product environments, slow start will take a relatively long time to achieve full link utilization. For error recovery, TCP uses the go-back-*N* ARQ strategy, also known to perform poorly in large bandwidth–delay product environments. Finally, asymmetric links may result in poor performance if the low-rate return path becomes congested with acknowledgments before the high-rate forward path is fully utilized.

There are several ways in which the problem of poor TCP performance in LMSC and other wireless systems may be addressed. One solution is to simply use an alternate protocol, one optimized for the satellite environment. Several such protocols have been proposed, including the Satellite Transport Protocol (STP) [37] and the Wireless Transmission Control Protocol (WTCP) [38]. Another approach is to modify or extend TCP to improve its performance over satellite links. Efforts in this area include the extensions proposed by Jacobson et al. [39], the selective acknowledgments (SACK) extension proposed by Mathis et al. [40], and the space communications protocol standards (SCPS) [41]. One serious drawback to TCP replacement and extensions is that in order to achieve the performance gains, all participating hosts must comply with the modification, thus negating the value of TCP's high penetration as a standard transport-layer protocol. Another approach is to employ TCP splitting. The idea behind TCP splitting is that the end-to-end TCP connection between two users is split into three segments. The first segment consists of a (presumably wired) TCP connection between the first user and a satellite terminal, the second segment consists of the satellite link, over which a protocol optimized for this environment is used, and the third link (also assumed to be wired) is a TCP connection between the receiving terminal and the second user. The splitting is done in such a way that it is transparent to the end users. In other words, from their perspective, an end-to-end TCP connection exists between them. However, because TCP is not actually used over the satellite link, the inefficiencies associated with this practice are not experienced. Of course, additional complexity must be introduced to perform the splitting, but this complexity is confined to the satellite terminals, and no modifications to end-user equipment (i.e., host computers) are required. Stadler et al. [36] and Mineweaser et al. [42] proposed and evaluated a TCP splitting mechanism, referred to as the *wireless IP suite enhancer* (WISE). In Fig. 11 the performance of WISE as a function of round-trip time is contrasted with that of TCP and TCP with the SACK extensions. In the figure, the transfer time of a 256-kilobyte (KB) file as a function of end-to-end round trip time is presented as an average over 20 simulation runs. The channel rate was taken to be 128 kbps and the average BER was 10^{-5}. In the WISE simulations, the protocol used over the satellite portion of the link is the Lincoln Laboratory Link Layer (LLLL) protocol [43]. Because this protocol resides at the link layer, it is also possible to use it directly with TCP (i.e., no splitting) to achieve some performance gain. With this approach, LLLL conditions the underlying satellite link according to the requirements of TCP. The performance of this scheme is also characterized in the figure. Note that WISE delivers nearly uniform performance regardless of round-trip time, while the other approaches are more sensitive to latency. Similar results can be achieved as a function of BER.

7. CONCLUSIONS

A brief description of several LMSC systems, their characteristics, and technical issues surrounding their

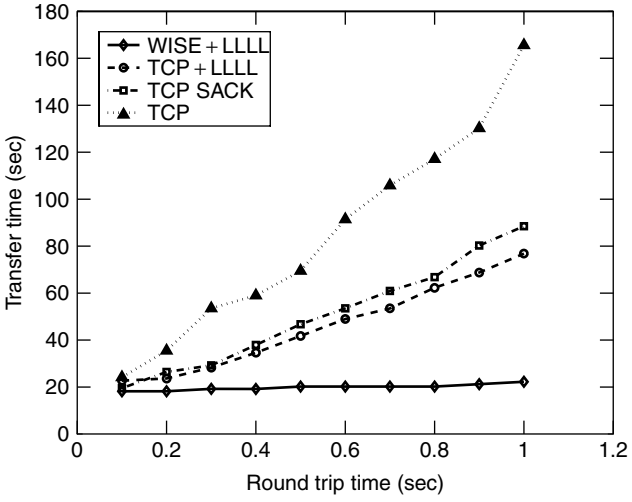


Figure 11. Average file transfer time as a function of round-trip time for several protocols.

implementation and operation has been presented. Major topics of discussion include the following. The LMSC propagation channel, including the effects of random noise, path loss, multipath fading, shadowing, attenuation due to antenna mispointing, was investigated. Also, error mitigation strategies such as FEC coding, ARQ protocols, and diversity techniques were examined. Multiple access schemes, including fixed assignment, random access, demand assignment, and adaptive assignment protocols, were also discussed. Finally, network aspects were covered, including a brief discussion of the similarities and differences in LMSC and terrestrial cellular voice networks, and an examination of the performance of TCP in networks that include LMSC links. Where appropriate, references were cited so that interested readers can pursue these topics in further detail.

BIOGRAPHY

Jeff Schodorf received his B.S.E.E., M.S.E.E., and Ph.D. in 1991, 1994, and 1996, respectively, all from the Georgia Institute of Technology. Since 1996 he has been a member of the Technical Staff in the Tactical Communication Systems Group at MIT Lincoln Laboratory. His research at the Laboratory has covered a variety of topics, including reduced complexity demodulation and decoding, satellite multiple access, and channel models and error control protocols for the land mobile satellite channel.

BIBLIOGRAPHY

1. T. T. Ha, *Digital Satellite Communications*, McGraw Hill, New York, 1990.
2. E. Lutz, M. Werner, and A. Jahn, *Satellite Systems for Personal and Broadband Communications*, Springer, 2000.
3. R. E. Sheriff and Y. F. Hu, *Mobile Satellite Communication Networks*, Wiley, New York, 2001.
4. D. Parsons, *The Mobile Radio Propagation Channel*, Halsted Press, New York, 1992.
5. J. G. Proakis, *Digital Communications*, McGraw-Hill, New York, 1989.
6. F. Xiong, *Digital Modulation Techniqes*, Artech House, Boston, 2000.
7. P. K. Karmakar et al., Radiometric measurements of rain attenuation at 22.2 and 31.4 GHz over Calcutta, *Int. J. Infrared and Millimeters Waves* 493–501 (1998).
8. S. Poonam and T. K. Bandopadhyaya, Rain rate statistics and fade distribution of millimeter waves in Indian continents, *Int. J. Infrared and Millimeters Waves* 503–509 (1998).
9. W. C. Jakes, *Microwave Mobile Communications*, IEEE Press, Piscataway, NJ, 1993.
10. E. Lutz et al., The land mobile satellite communications channel — recording, statistics and channel model, *IEEE Trans. Vehic. Technol.* 375–386 (May 1991).
11. M. Rice et al., K-band land-mobile satellite characterization using ACTS, *Int. J. Sat. Commun.* 283–296 (Jan. 1996).
12. M. Rice and B. Humphreys, Statistical models for the ACTS K-band land mobile satellite channel, *Proc. IEEE Vehicular Technology Conf.*, 1997.
13. M. Rice and B. Humphreys, A new model for the ACTS land mobile satellite channel, *Proc. Int. Mobile Satellite Conf.*, 1997.
14. W. J. Vogel and J. Goldhirsh, Multipath fading at L band for low elevation angle, land mobile satellite scenarios, *IEEE J. Select. Areas Commun.* 197–204 (Feb. 1995).
15. M. G. Shayesteh and A. Aghamohammadi, On the error probability of linearly modulated signals on frequency-flat Ricean, Rayleigh, and AWGN channels, *IEEE Trans. Commun.* 1454–1466 (Feb. 1995).
16. A. J. Simmons, *EHF Propagation through Foliage*, MIT Lincoln Laboratory Technical Report TR-594 1981.
17. C. Loo, A statistical model for a land mobile satellite link, *IEEE Trans. Vehic. Technol.* 122–127 (Aug. 1985).
18. G. E. Corazzo and F. Vatalaro, A statistical model for land mobile satellite channels and its application to nongeostationary orbit systems, *IEEE Trans. Vehic. Technol.* 738–742 (Aug. 1994).
19. F. P. Fontan et al., Statistical modeling of the LMS channel, *IEEE Trans. Vehic. Technol.* 1549–1567 (Nov. 2001).
20. A. C. Densmore and V. Jamnejad, A satellite tracking K- and Ka-band mobile vehicle antenna system, *IEEE Trans. Vehic. Technol.* 502–513 (Nov. 1993).
21. A. Densmore et al., K- and Ka- band land mobile satellite-tracking reflector antenna system for the NASA ACTS mobile terminal, *Proc. Int. Mobile Satellite Conf.*, 1993.
22. M. Rice, B. J. Mott, and K. D. Wise, A pointing error analysis of the ACTS mobile terminal, *Proc. Int. Mobile Satellite Conf.*, 1997.
23. J. B. Schodorf, A probabilistic mispointing analysis for land mobile satellite communications systems with directive antennas, *Proc. IEEE Vehicular Technology Conf.*, 2001.
24. S. Lin and D. J. Costello, *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
25. C. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error correcting coding and decoding: Turbo codes, *Proc. IEEE Int. Conf. Communication*, 1993.
26. B. Sklar, A primer on turbo code concepts, *IEEE Commun. Mag.* 94–102 (Dec. 1997).
27. J. B. Schodorf, Error control for Ka-band land mobile satellite communications systems, *Proc. IEEE Vehicular Technology Conf.*, 2000.
28. J. Schindall, Concept and implementation of the Globalstar mobile satellite system, *Proc. Intl. Mobile Satellite Conf.*, 1995.
29. R. Akturan and W. J. Vogel, Path diversity for LEO satellite-PCS in the urban environment, *IEEE Trans. Antennas Propag.* 1107–1116 (July 1997).
30. T. Nguyen and T. Suda, Survey and evaluation of multiple access protocols in multimedia satellite networks, *Proc. Southeastcon*, 1990, 408–412.
31. A. J. Viterbi, A perspective on the evolution of multiple access satellite communication, *IEEE J. Select. Areas Commun.* 980–983 (Aug. 1992).
32. D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed., Prentice-Hall, Upper Saddle River, NJ, 1992.
33. L. G. Roberts, Dynamic allocation of satellite capacity through packet reservation, *Proc. Nat. Computer Conf., AFIPS Conf.*, 1973, pp. 711–716.

34. L. Kleinrock and Y. Yemini, An optimal adaptive scheme for multiple access broadcast communication, *Proc. IEEE Int. Conf. Communication*, 1978, pp. 7.2.1–7.2.5.

35. I. M. Jacobs, R. Binder, and E. V. Hoversten, General purpose packet satellite networks, *Proc. IEEE* 1448–1467 (Nov. 1978).

36. J. S. Stadler, J. Gelman, and J. Howard, Performance enhancements for TCP/IP on wireless links, *Proc. Virginia Tech/MPRG Symp. Wireless Personal Communications*, 1999.

37. T. R. Henderson and R. H. Katz, Transport protocol for internet-compatible satellite networks, *IEEE J. Select. Areas Commun.* 326–344 (Feb. 1999).

38. P. Sinha et al., WTCP: A reliable transport protocol for wireless wide-area networks, *Wireless Networks* 301–316 (2002).

39. V. Jacobson, R. Braden, and D. Borman, *TCP Extensions for High Performance*, IETF, RFC 1323, May 1992.

40. M. Mathis et al., TCP selective acknowledgment options, IETF, RFC 2018, Oct. 1996.

41. R. C. Durst, G. J. Miller, and E. J. Travis, TCP extensions for space communications, *Wireless Networks* 389–403 (1997).

42. J. L. Mineweaser et al., Improving TCP/IP performance for the land mobile satellite channel, *Proc. IEEE Military Communication Conf.*, 2001.

43. J. S. Stadler, A link layer protocol for efficient transmission of TCP/IP via satellite, *Proc. IEEE Military Communication Conf.*, 1997.

LEAKY-WAVE ANTENNAS

Fabrizio Frezza
Alessandro Galli
Paolo Lampariello
"La Sapienza" University of Rome
Roma, Italy

1. INTRODUCTION

1.1. Definition

Leaky-wave antennas (LWAs) constitute a type of radiators whose behavior can be described by an electromagnetic wave ("leaky wave") that propagates in guiding structures that do not completely confine the field, thus allowing a continuous loss of power towards the external environment ("leakage").

According to the IEEE Standard 145-1993, a leaky-wave antenna is "an antenna that couples power in small increments per unit length either continuously or discretely, from a traveling wave structure to free space."

1.2. General Properties and Applications

LWAs [1] belong to the class of traveling-wave line antennas, for which the illumination is produced by a wave that propagates along a guiding structure [2]. If compared with the wavelength, a LWA is "long" in the propagation direction z, while its cross section is usually

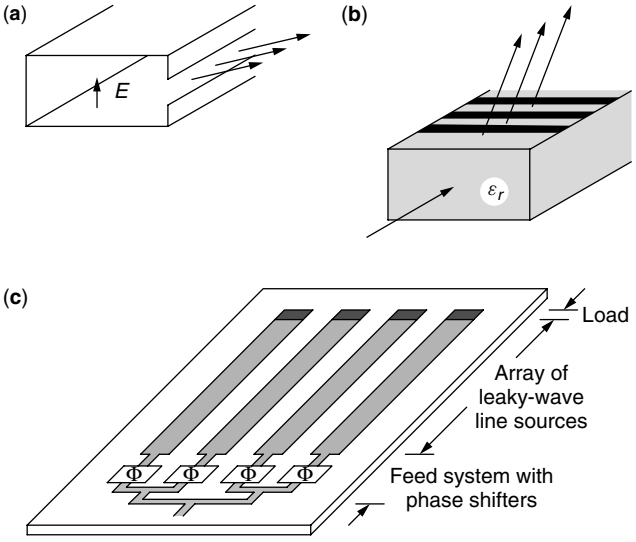


Figure 1. Basic structures of leaky-wave antennas (LWAs): (**a**) uniform LWAs: — geometry derivable by a partially open metallic waveguide; (**b**) periodic LWAs: — geometry derivable by a strip-loaded dielectric-rod waveguide; (**c**) topology of LWA arrays.

of the order of the wavelength (see the reference examples of Fig. 1a,b).

LWAs radiate along their lengths and in general are excited from one input of the open guiding structure with a traveling wave that propagates mainly in one longitudinal direction (e.g., $+z$) and is attenuated as a result of the power leakage toward the exterior region, thus leaving a negligible field at the end termination of the guide. In a harmonic regime (with an $\exp(j\omega t)$ time dependence), this wave is characterized by a complex propagation constant of the type $k_z = \beta_z - j\alpha_z$ [3,4], where β_z is the "phase constant" and α_z is the "attenuation constant" of the leaky wave (when only power loss due to radiation is taken into account, α_z is also said "leakage constant").

Usually the radiation pattern of a single LWA has a typical "fan" shape; in the elevation (or zenith) plane a narrow beam is achievable with the pointing direction that varies by frequency, while in the cross (or azimuth) plane the beam is usually wider in connection with the characteristics of a more reduced transverse aperture. Depending on the desired application, a suitable longitudinal variation of the aperture distribution, usually reached by modulating geometric parameters ("tapering"), allows a good control of the radiation pattern (sidelobe behavior, etc.). In some cases, in order to obtain a beam shaping or a physical matching with mounting curved surfaces, LWAs can be designed with certain amounts of curvature along their lengths [5].

The scanning properties in the elevation plane (pointing angle variable with the frequency) are related to the type of waveguide employed, which can be of either "uniform" (Fig. 1a) or "periodic" type (Fig. 1b) [1,2]. LWAs derived by waveguides that are longitudinally uniform (i.e., the structure maintains continuously the same transverse geometry) typically allow the angular scanning in one quadrant, from around broadside toward one endfire

(the "forward" one, which is concordant with the wave propagation direction). LWAs derived by waveguides that are longitudinally periodic (i.e., where the structure is periodically loaded with proper discontinuities, at intervals that are usually short with respect to the wavelength) allow a wider angular scanning in both the forward and backward quadrants. However, due to different causes, limitations in such scanning ranges generally exist for both the types of structures. A scan range in both quadrants may also be accomplished by using anisotropic media.

When a "pencil beam" is aimed with a possible two-dimensional (2D) scanning in both elevation and cross-planes (zenith and azimuth), a phased array of juxtaposed LWAs may be employed, thus enlarging the equivalent aperture also transversely [6,7] (Fig. 1c). LWA arrays are therefore constituted by a linear configuration of sources (i.e., one-dimensional elements), instead of the planar ones of standard arrays (i.e., two-dimensional elements). For LWA arrays a pointed-beam scanning is achievable by varying both the frequency for the elevation plane and the phase shift for the cross-plane.

Since LWAs are derived by partially open waveguides, they present a number of distinctive features as radiators: handling of high-power amounts, particularly for structures derivable by closed metallic waveguides; reduction of bulk problems, due to the usually small profiles in the cross sections; capability of designing a wide variety of aperture distributions and consequent flexibility for the beamshaping; possible use as wideband radiators, allowing large angular scanning by varying frequency (instead of using mechanical or other electronic means); achievement of very narrow beams with good polarization purity; and simplicity of feeding and economy for 2D scannable pencil-beam arrays (reduced number of phase shifters).

LWAs are used mainly in the microwave and millimeter wave regions. The first studies on LWAs were presented during the 1940s, basically for aerospace applications (radar, etc.); since then, a very wide number of different solutions for LWAs has been proposed in connection with changing requirements and constraints. Also the applicability of this type of antennas has been widened, involving various problems of traffic control, remote sensing, wireless communications, and so forth [8].

2. PRINCIPLES OF OPERATION

2.1. Leaky Waves in Open Structures

A leaky wave [3,4] has a complex longitudinal wavenumber k_z that can be derived by solving, as a function of the physical parameters (frequency and geometry of an open waveguiding structure), the characteristic equation (or dispersion relation), which is of the general type:

$$D(k_z, k_0) = 0 \tag{1}$$

where $k_0 = \omega(\mu_0\varepsilon_0)^{1/2}$ is the vacuum wavenumber.

As is well known, for lossless closed waveguides the dispersion relation (1) generally presents an infinite discrete set of eigensolutions giving the "guided modes," which individually satisfy all the relevant boundary conditions. Any field excited by a source in a closed guide can be expanded in terms of the complete set of the infinite discrete eigensolutions derived by Eq. (1). In conventional guides, the longitudinal wavenumbers k_z are either real [propagating waves above their cutoff, with $k_z = \beta_z < k = k_0(\varepsilon_r)^{1/2}$] or imaginary (attenuating waves below their cutoff, with $k_z = -j\alpha_z$).

In lossless open waveguides (e.g., dielectric guides), instead, only a finite number of propagating modes can exist as eigensolutions of Eq. (1) satisfying all the boundary conditions (particularly the radiation condition); these are the so-called bound "surface waves" (each one exists only above its cutoff, with $k_z = \beta_z > k_0$). In addition to this, for a complete representation of the field that is no longer confined to a closed section, a "continuous spectrum" of modes must be introduced to describe the radiated field as an integral contribution in terms of a set of plane waves having a continuous range of wavenumbers (e.g., such that $0 < k_z = \beta_z < k_0$ and $-j\infty < k_z = -j\alpha_z < 0$). Any field excited by a source in an open guide can therefore be expanded by means of a "spectral representation," that is, in terms of a finite set of guided modes and an integral contribution of the continuous spectrum.

On the other side, it is seen that the characteristic equation (1) for open guides presents additional discrete solutions that are "nonspectral," since they correspond to fields that violate the radiation condition (they attenuate along the propagation direction but exponentially increase in a transverse direction away from the structure) and are not included in the spectral representation of the field. In an open lossless structure the leaky-wave solutions that are of the type $k_z = \beta_z - j\alpha_z$ describe power flowing away from the structure.

In many practical circumstances, for describing the radiative effects of the open structures in the presence of a source, the evaluation of the field through the "spectral representation" (i.e., including the integral contribution of the continuous spectrum) can be very difficult and cumbersome to quantify. It is seen that the radiation field can be evaluated accurately in much a simpler fashion by considering just the contribution due to the presence of one complex mode, that is, a leaky wave, which can therefore be viewed as a simple rephrasing of the continuous spectrum. In fact, it is seen that in practical cases the remaining part of the continuous spectrum (viz., the "space wave" or "residual wave") is able to give negligible contributions to the description of the LWA's radiation.

It can be seen that, when properly excited by a source at a finite section, a leaky wave, even though nonspectral, assumes its physical validity within an angular sector close to the equivalent aperture of the open guiding structure, and the relevant field distribution is able to furnish a fundamental contribution to evaluation of the near field. Since the relevant far field is achieved by a simple Fourier transform of the field on the aperture, a leaky wave can definitively furnish a highly convergent and efficient quantification of the radiation of LWAs, as an extremely advantageous alternative to a continuous spectrum evaluation.

2.2. Characterization of Leaky-Wave Antennas

LWAs present the advantage of a rather simple characterization of their basic properties, with consequent straightforward approaches for the analysis and synthesis. The basic knowledge is reduced to the evaluation of a dominant complex eigensolution $k_z = \beta_z - j\alpha_z$ that can be supported and strongly excited in a specific open structure.

The characteristic behaviors of the real and imaginary parts of the longitudinal wavenumber of a leaky wave are presented in Fig. 2; specifically the dispersion behaviors of the normalized parameters β_z/k_0 and α_z/k_0 versus frequency f. The radiation region of LW structures lies approximately inside the frequency range where the wave becomes fast ($\beta_z/k_0 < 1$) and power can therefore leak out from the guiding structure toward the outside air region in the typical form of a TEM-like mode; in fact, $\beta_z/k_0 < 1$ is in general the so-called condition for leakage of a complex wave that can radiate in an external air region.

The valid frequency range for LWA applications is actually where, as the frequency decreases, β_z/k_0 diminishes monotonically from unity toward rather low values; in this region, to have an efficient directive beam, α_z/k_0 should assume rather limited values (e.g., typically α_z/k_0 varies from about 10^{-1} to 10^{-3}). As frequency decreases further, a sudden rise of α_z/k_0 is generally found, which describes the predominance of reactive phenomena instead of radiative ones, while β_z/k_0 can present a flat zone with approximately constant low values before showing a further steep rise as frequency goes to zero: in these ranges, radiative effects can no longer be represented by the leaky wave and the structures usually cannot work well as radiators [3,4,7].

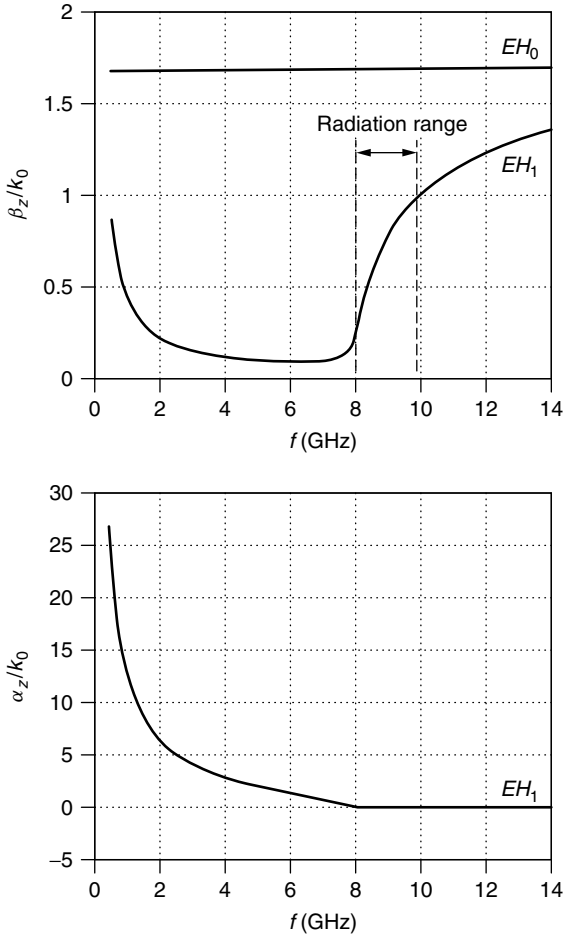


Figure 2. Typical dispersion behavior of the leaky-mode complex wavenumber (normalized phase β_z/k_0 and leakage α_z/k_0 constants vs. frequency f) for an open planar waveguide (microstrip).

It is worth noting here that in open planar structures a different type of leakage can occur as well, which is associable to "surface waves" (that are TE- or TM-like modes) propagating in the substrates [7], instead of the standard "space wave" (TEM-like mode) that carries out power in the outside air region; while the latter is able to account for useful contributions to far-field radiation in LWA's applications, the former usually describes power that leaks out transversely in a layered structure and accounts for loss and interference effects in the planar circuits.

2.3. Evaluation of the Leaky-Wave Phase and Leakage Constants

The evaluation of the complex eigensolutions for nonclosed waveguides depends on the physical parameters involved (frequency and geometry) and is generally achievable with numerical methods. Among them, the *transverse resonance technique* (TRT) [9,10] is one of the most efficient approaches for either rigorous or approximate (according to the antenna topology) evaluations. It first requires the introduction of a suitable equivalent transmission-line network, which describes the transverse geometry of the structure. Then, a numerically solvable transcendental equation in terms of transverse eigenvalues k_t and of physical parameters is usually achievable by imposing a resonance condition for the equivalent circuit. The complex eigenvalue k_z is derived by the additional link to the longitudinal problem given by a separation condition for the eigenvalues (e.g., in air: $k_0^2 = \omega^2\mu_0\varepsilon_0 = k_t^2 + k_z^2$). Where the separation condition holds rigorously also for the variables in the transverse plane (e.g., $k_t^2 = k_x^2 + k_y^2$), TRT in general gives exactly the characteristic equation of the geometry. An example is given in the next paragraph.

When separation of variables does not strictly hold, other numerical methods can nevertheless be employed to accurately determine the complex eigensolutions of the involved open waveguides. The most appropriate choice depends on several factors related to the computational features of the methods, the geometry of the open-type structures, and so on [9,10]. Among the various possible approaches, integral equation techniques can work particularly well. As is known, in particular spectral domain approaches appear well suited for the derivation of the eigensolutions in structures of printed type [9].

2.4. Interpretation of the Behavior of a Leaky-Wave Antenna

As stated above, LWAs are described by a fast wave that propagates on an equivalent aperture losing power

toward free space with a leakage amount that is usually rather limited to allow a sufficiently directive beam. The simplest LWA geometry for this purpose is derivable by a closed metallic waveguide in which a suitable "small" aperture is introduced longitudinally in order to get a continuous power loss along its length, as shown in Fig. 3a for a rectangular guide with a slit cut on a sidewall. This structure, besides having a historical importance as the first proposed LWA in 1940 [1,2], can be taken as a reference structure for explaining the basic behavior of LWA's in terms of a waveguide description.

For such a structure, a leaky wave can be considered as excited by a standard incident mode for the closed rectangular waveguide, that is the dominant TE_{10}, which travels in the $+z$ direction with a known phase constant β_{0z} for a fixed choice of the physical parameters (geometry and frequency). For a sufficiently small geometry perturbation due to the slit, the phase constant is changed just slightly to a value represented by β_z, and a "low" leakage rate

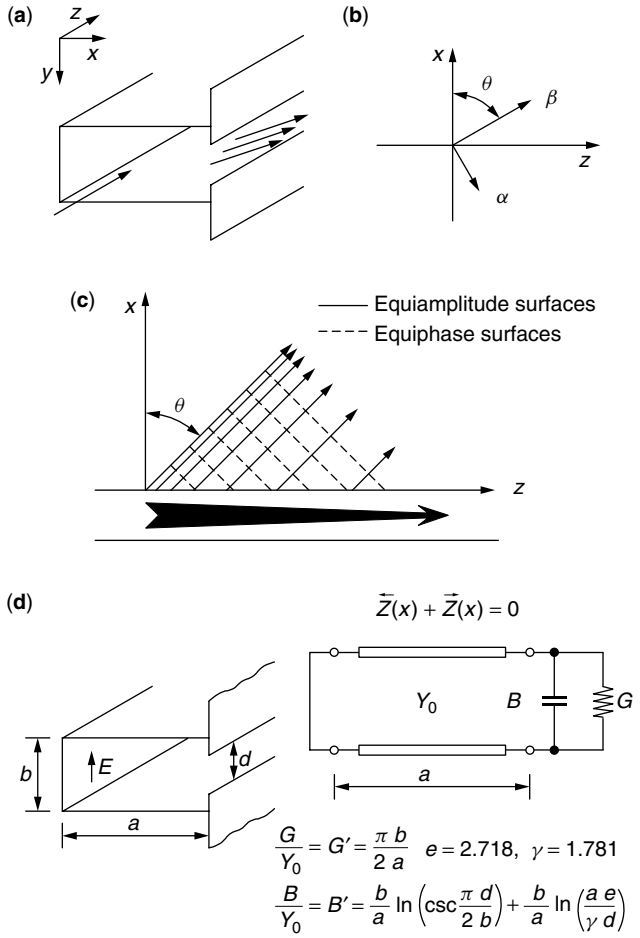


Figure 3. LWA derived by a sidewall slit rectangular waveguide: (**a**) geometry of the structure; (**b**) nature of the propagation vector of the inhomogeneous plane leaky wave (phase and attenuation vectors); (**c**) equiphase and equiamplitude planes of the leaky wave with the relevant leakage phenomenon along the guide; (**d**) equivalent transverse resonance network, resonance conditions, and network parameters for the numerical evaluation of the leaky-wave complex wavenumbers as a function of the physical parameters involved.

α_z originates, too, which as mentioned accounts for the longitudinal attenuation due to the field that is no longer confined and flows also in the outside region; the propagating field inside the waveguide and in the proximity of its aperture is therefore described by the complex longitudinal wavenumber $k_z = \beta_z - j\alpha_z$, whose quantification depends on the physical parameters.

In this case the leakage phenomenon is assumed along $+z$ ($\beta_z > 0$ and $\alpha_z > 0$), and by supposing that the vertical field variations are almost negligible ($k_y \cong 0$), it is easily seen that, from the general separation condition for waveguides ($k_0^2 = \omega^2\mu_0\varepsilon_0 = k_t^2 + k_z^2 \cong k_x^2 + k_z^2$), the horizontal wavenumber is also complex:

$$k_x = \beta_x - j\alpha_x, \tag{2}$$

with $\beta_x > 0$ and $\alpha_x < 0$ since it results $\beta_x\alpha_x = -\beta_z\alpha_z$. Therefore a plane wave of inhomogeneous type exists, having a complex propagation vector $\mathbf{k}$ of the type.

$$\begin{aligned} \mathbf{k} &= \boldsymbol{\beta} - j\boldsymbol{\alpha} \\ \boldsymbol{\beta} &= \beta_x\mathbf{x}_0 + \beta_z\mathbf{z}_0 \\ \boldsymbol{\alpha} &= \alpha_x\mathbf{x}_0 + \alpha_z\mathbf{z}_0 \end{aligned} \tag{3}$$

with the phase vector $\boldsymbol{\beta}$ directed at an angle that describes the outgoing of power from the guide to the external, and the attenuation (leakage) vector $\boldsymbol{\alpha}$ that is perpendicular to $\boldsymbol{\beta}$, and represents attenuation along z and amplification along x. Consequently, the field has a spatial dependence of the type

$$\exp[-j(\beta_x x + \beta_z z)]\exp[|\alpha_x|x - \alpha_z z] \tag{4}$$

Therefore, this plane wave travels at an angle $\theta = \sin^{-1}(\beta_z/|\boldsymbol{\beta}|)$ with respect to broadside carrying out power, and its amplitude is transversely increasing as expected in a leaky wave. It should be noted that the direction angle θ of the leaky wave is usually expressed under the approximate form: $\theta \cong \sin^{-1}(\beta_z/k_0)$, since in general the leakage constant is numerically negligible with respect to the phase constant. The nature of the propagation vector is sketched in Fig. 3b, while the distribution of equiphase and equiamplitude surfaces with the decreasing power flow along the guide is represented in Fig. 3c. It should be recalled that, even though the leaky wave has a nonspectral nature, the field generated from a source located at a finite distance along z still satisfies the radiation condition, since the field increases transversely only in a limited sector given by angles less than the θ value describing the direction of power leakage [3,4].

A quantitative description of this LWA is easily achieved with a simple analysis of the complex eigenvalue derivable as a modification of the dominant mode by employing a TRT [1,2,11]. To this aim, it is required a characterization of the slit aperture in the side wall as a circuit element in the equivalent transmission line. For the quantification of such discontinuities, a great deal of work was developed since the 1950s, basically through variational methods [2,7,9–12]. The description of the radiative and reactive effects of the slit in the side wall of the rectangular guide can be represented by a lumped

element (e.g., an admittance $Y_R = G_R + jB_R$) as a function of geometry and frequency. The transverse network is reported in Fig. 3d. The solution of the relevant resonance equation in the complex plane for the perturbed dominant mode describes the leaky-wave behavior.

3. DESIGN PROCEDURES

3.1. Basic Radiation Features

The basic design principles of LWAs are generally derivable from the knowledge of the desired beam width and of the pointing direction. In LWAs these quantities can be linked in a straightforward way to the complex longitudinal wavenumber.

In fact, the beam maximum direction θ_M is, as seen before, related mainly to the normalized phase constant, according to the simple relationship

$$\sin\theta_M \cong \frac{\beta_z}{k_0} \tag{5}$$

Since β_z has a dispersive behavior as is typical of waveguiding structures, a change in the frequency yields a scanning of the beam: typically, as the frequency is increased from the cutoff, the pointing angle varies its direction from around the broadside ($\theta_M = 0^\circ$), toward the forward endfire ($\theta_M = 90^\circ$).

In regard to the beamwidth, we recall that the leakage constant α_z quantifies the rate of power loss along the line due to the radiation, thus influencing primarily the effective dimension of the equivalent aperture for the line source: in fact, the more α_z increases, the more the actual illumination length reduces (and the less the beamwidth is focused).

A basic link between the leakage constant and the antenna length L derives from the specification of the radiation efficiency η, expressible in LWAs as $\eta = [P(0) - P(L)]/P(0)$, where $P(0)$ is the input power delivered to the structure and $P(L)$ is the output power left at the end termination. The link between efficiency, leakage rate, and length is generally dependent on the desired radiation pattern and therefore on the aperture distribution: referring to a uniform-section LWA, where α_z is independent of z, this results in $\eta = 1 - \exp(-2\alpha_z L)$. It should also be noted that, for narrowbeam applications, a very high increase in efficiency should require an extreme prolongation of the line source; actually, in LWAs it is typical to radiate around 90% or at most 95% of the input power, where the remaining power at the end termination is absorbed by a matched load to avoid a backlobe of radiation due to the reflected wave.

Once the efficiency is chosen, a fixed link therefore exists between the relative length in terms of wavelengths L/λ_0 and the normalized leakage constant α_z/k_0. For a uniform-section LWA, an inverse proportionality relationship between L and α_z is found of the type

$$\frac{L}{\lambda_0} \cong \frac{c}{\alpha_z/k_0}$$
$$c = \left(\frac{1}{4\pi}\right)\ln\left(\frac{1}{1-\eta}\right) \tag{6}$$

where c is related to the value of the desired efficiency (e.g., for 90% of efficiency it is $c = 0.185$). For a nonuniform section, since α_z depends on z, the link between efficiency, length, and leakage rate is related to the chosen illumination and is more complicated.

In order to achieve narrow beams in the elevation angle, the effective longitudinal aperture has to be sufficiently wide (usually several wavelengths), and this implies a rather low leakage rate. The half-power (−3-dB) beamwidth $\Delta\theta$ is directly linkable to the normalized antenna length L/λ_0 through an approximate relationship, which takes into account also the contribution of the scan angle [1]:

$$\Delta\theta \cong a \Big/ \left(\frac{L}{\lambda_0}\cos\theta_M\right) \quad \text{(rad)} \tag{7}$$

where the proportionality factor a is dependent on the aperture distribution; it has the most reduced value for a constant aperture distribution ($a \cong 0.88$) and increases for tapered distributions (typically, more than unity) [1]. From the previous expression, it is seen that, since $\cos\theta_M \cong k_t/k_0$, the beamwidth is also expressible as $\Delta\theta \cong 2\pi/(k_t L)$. This means that the beam width is, as a first approximation, practically constant when the beam is scanned away from broadside by varying the frequency for air-filled LWAs (where k_t is independent of frequency), while it changes for dielectric-filled LWAs (where k_t depends on frequency).

The effective aperture is anyway reduced for a fixed antenna length as the beam approaches endfire (where the previous expression becomes not accurate), and $\Delta\theta$ anyway tends in practice to enlarge. It can be seen that for an ideal semiinfinite uniform structure, that is an antenna aperture from $z = 0$ to $z = L \to \infty$, the beam width is determined by the leakage rate only, since in this case it can be found that $\Delta\theta \cong 2\alpha_z/k_t$. Moreover, in this situation the radiation pattern depends only on β_z and α_z and does not present sidelobes:

$$R(\theta) \approx \frac{\cos^2\theta}{(\alpha/k_0)^2 + (\beta/k_0 - \sin\theta)^2} \tag{8}$$

For finite antenna lengths, sidelobes are produced and the expression for $R(\theta)$ is more involved. In general the specifications on the sidelobe level are related to the choice of the aperture distribution, whose Fourier transform allows the derivation of the radiation pattern.

3.2. Scanning Properties

It is seen that the beams for LWAs derived by partially open air-filled metallic waveguides scan in theory an angular region from around the broadside ($\beta_z/k_0 \cong 0$) towards one endfire ($\beta_z/k_0 \cong 1$).

In practice, around broadside the structure works near the cutoff region of the closed waveguide where reactive effects are increasingly important. The leaky-wave values for β_z/k_0 cannot anyway be extremely low and at the same time α_z/k_0 tends to increase too much, adversely affecting the possibility of focusing radiation at broadside.

Concerning the behavior at endfire it is seen that, since β_z/k_0 tends to unity asymptotically as the frequency

increases, in the unimodal range (where these structures are usually employed) the beam cannot reach so closely the endfire radiation in an air-filled LWA. A way of improving the angular scanning is to fill these structures with dielectric materials. Thus, since in this case the normalized phase constant approaches the square root of the relative permittivity as the frequency is increased ($\beta_z/k_0 \to \varepsilon_r^{1/2}$), the $\beta_z/k_0 = 1$ value can actually be approached in a much more restricted frequency range. It should anyway be noted that for such dielectric-filled structures the beam width may change strongly as a function of frequency and therefore as the pointing angle varies [see comments on Eq. (7)].

Moreover, it should be noted that in many leaky structures (such as the dielectric and printed ones), as the frequency is increased, the leaky-mode solution changes into a guided-mode solution through a complicated "transition region" [13,14]; in this frequency range, also called "spectral gap," the contribution of the leaky wave to the field tends progressively to decrease, and generally the structure does not work well as a LWA.

As stated above, while the uniform LWAs usually radiate only in the forward quadrant, with the limits specified above, the LWAs derived from periodically modulated slow-wave guides can start to radiate from the backward endfire in the lower frequency range.

The design principles of periodic LWAs are in most part similar to those of uniform LWAs [1,2]. The main difference lies in the characterization of the fast wave that is now associated to a Floquet's space harmonic of the periodic guide [1,2,14,15]. One can see that if a uniform guide is considered whose operating mode is slow ($\beta_z/k_0 > 1$, e.g., a dielectric waveguide), and a longitudinally-periodic discontinuity is properly added (e.g., an array of metal strips or notches, placed at suitable distances p), such periodicity furnishes a field expressible in an infinite number of space harmonics ($\beta_{zn}p = \beta_{z0}p + 2n\pi$), where β_{z0} is the phase constant of the fundamental harmonic, which is slightly varied with respect to the original value β_z of the unperturbed guide. With proper choices of the physical parameters, it is in general possible to make only one harmonic fast (typically, the $n = -1$), so that it can radiate as a leaky wave (presence of an additional attenuation constant α_z).

In this case, the phase constant of this fast harmonic can assume both positive and negative values ($-1 < \beta_z/k_0 < 1$), as a function of the parameters involved; in particular, as frequency is increased, the beam starts to radiate from backward endfire toward the broadside. In general, also periodic LWA's have difficulties in working well in the broadside region, since usually for periodic structures there exists an "open stopband" [14], where the attenuation constant rapidly increases, resulting in a widening beamwidth.

As the frequency is further increased after broadside, the beam is then scanned also in the forward quadrant. In periodic LWAs, depending on the choice of the design parameters, additional limitations in the forward scanning behavior could exist when also a second harmonic starts to radiate before the first harmonic reaches its endfire, thus limiting the single-beam scanning range [1,14].

3.3. Leaky-Wave Arrays for Pencil-Beam Radiation

If an increase of directivity in the cross-plane is desired, a simple improvement for LWAs based on long radiating slots can be achieved by a physical enlargement of the transverse aperture (e.g., with a flared transition to enlarge the effective cross-aperture). As said before, a more efficient way to increase directivity in the cross-plane is to use a number of radiators placed side by side at suitable lateral distances, thus constituting a linear array; it is then possible to achieve radiation with a focused pencil beam. In addition, if properly phased, these arrays of LWAs allow a 2D scanning of the beam: in the elevation plane, as is typical for LWAs, the scanning is achievable by varying the frequency, while in the cross-plane the scanning is achievable with phase shifters that vary the phase difference among the single line sources. As noted, in LWAs only a unidimensional number of phase shifters is therefore necessary, with particular structural simplicity and economic advantage if compared to all the usual radiators requiring a 2D number of shifters for the scanning. Additional desirable features of such arrays are in general the absence of grating lobes and of blind spots, and good polarization properties.

For analysis of such LW arrays, an efficient method is that one based on the "unit cell" approach [6,7]. In this way, it is possible to derive the behavior of the global structure by referring to a single radiator taking into account the mutual effects due to the presence of all the others. In the equivalent network this is achievable by changing only the description of the radiation termination for a periodic-type array environment (infinite number of linear elements): in particular, an "active admittance" can be quantified, which describes the external radiating region as a function of the geometry and of the scan angle. More sophisticated techniques also allow accurate analyses of arrays by taking into account the mutual couplings for a finite number of elements [6].

3.4. Radiation Pattern Shaping

In the basic requirements of the radiation pattern, in addition to the specification for the maximum of the beam direction and for its half-power width, also the sidelobe behavior has a primary importance. In a general sense, it is desired to derive the properties of the source in connection with a desired radiation pattern. Since LWAs can be viewed as aperture antennas with a current distribution having a certain illumination $A(z)$, it is possible to obtain the far field through a standard relationship:

$$E(\theta) = G(\theta) \int_0^L |A(z')| e^{j\mathrm{Arg}[A(z')]} e^{jkz' \sin\theta} dz' \qquad (9)$$

The radiation pattern for E is expressed in terms of a Fourier transform of the line-source complex current distribution on the aperture multiplied by the pattern of the element current G (e.g., a magnetic dipole).

It is easily seen that if the LWA geometry is kept longitudinally constant, the amplitude distribution has always an exponential decay of the type: $\exp(-\alpha_z z)$. As is known, this behavior furnishes a quite poor radiation

pattern for the sidelobes that are rather high (around −13 dB). It therefore derives that, in conjunction with the choice of a fixed illumination function $A(z)$ giving a desired sidelobe behavior (e.g., cosine, square cosine, triangular, Taylor), the leakage rate has to be modulated along the main direction z of the line source: in practice this is achievable by properly modifying the cross section of the LWA structure along z, with the procedure usually known as "tapering." Considering that, for a smoothly tapered antenna, the power radiated per unit length from the antenna aperture is simply related to the aperture distribution [viz. $-dP(z)/dz = 2\alpha_z(z)P(z) = c|A(z)|^2$], a useful analytic expression for $\alpha_z(z)$ as a function of the amplitude $A(z)$, of the line-source length L, and of the efficiency η is obtainable [1,2,16]:

$$\alpha_z(z) = \frac{1}{2}\frac{|A(z)|^2}{\frac{1}{\eta}\int_0^L |A(z')|^2 dz' - \int_0^z |A(z')|^2 dz'} \tag{10}$$

From this equation it is also seen that the more the efficiency is desired high (around unity), the more α_z has to increase toward extremely high values around the terminal section (as mentioned, efficiency in common practice does not exceed 90–95%).

In general, in the tapering procedure the longitudinal modification of the geometry should be made in an appropriate way in order to affect only the leakage constant, taking into account that the phase constant should conversely be maintained the same (in pencil-beam applications, β_z should not depend on z in order to have the correct pointing angle for each elementary current contribution on the aperture).

The pattern shaping procedure requires therefore the knowledge of the phase and leakage constants as a function of the geometric and physical parameters of the chosen structure, and this is achievable, as stated, by finding the suitable complex eigensolution with numerical methods. Since the pattern shaping requires a proper α_z distribution with β_z constant, the procedure is strongly simplified if it is possible to find geometric parameters through which the leakage and phase constants are varied as independently as possible. This property is related to the topology characteristics of the waveguiding structure.

An example of tapering is sketched in Fig. 4 for a leaky structure, the so-called "stepped" LWA (Fig. 4a), proposed for high-performance applications with well-controlled radiation patterns [17] and additional general desirable features (increased geometrical flexibility, compactness, low profiles for aerospace applications, etc.).

In Fig. 4b the detailed behavior of the modulation in the height of the lateral steps is shown as a function of z for a desired illumination (cosine type). A first action only on the steps' unbalance, with their mean value kept constant (dashed profile), modifies appropriately the longitudinal distribution of the leakage constant, maintaining almost constant the phase constant. A second action is advisable to compensate the phase nonlinearity, which can give rather disturbing effects on the radiation patterns: in this topology it is possible to slightly vary the steps' mean value, with the previously fixed unbalance, to obtain the final valid profile (solid line). The relevant radiation patterns are then illustrated in Fig. 4c,d, for the single-shot and the double-shot tapering procedures, respectively; Fig. 4c is a rather "distorted" pattern related to the nonoptimized tapering (dashed profile), while Fig. 4d is a "correct" cosine-type pattern related to the optimized tapering (solid profile). The tapering procedure can be performed numerically in an easy way from a TRT network representation of the structure. The typical scanning behavior of these kinds of antennas is finally illustrated in Fig. 4e for the pointed beam variable by frequency.

4. FURTHER EXAMPLES OF SPECIFIC STRUCTURES

4.1. Partially Open Metallic Waveguides

One of the main drawbacks of the antenna shown in Fig. 1a is related to the leakage constant, which in general cannot be reduced below a certain limit. Reduced leakage amounts are achievable by slitting the top wall of the rectangular guide, decreasing the current modification due to the cut (Fig. 5a). By shifting the cut with respect to the central vertical plane, it is possible to modulate the leakage rate: investigations were also performed with tapered meander profiles for sidelobe control [18].

A way of improving the polarization purity in the basic geometry of a top-wall slitted rectangular guide is to use an aperture parallel-plate stub, able to reduce the contribution of the higher modes on the aperture, which are below cutoff, while the dominant leaky wave travels nonattenuated as a TEM-like mode at an angle [19] (Fig. 5b). Metal wide flanges, simulating an open half-space on the upper aperture, can increase the directivity of this type of LWA.

4.2. Printed Lines: Microstrip LWAs

The possibility of using LWAs also in printed circuitry has received interest that is probably destined to increase in the near future due to the wide use of planar technology for light, compact, and low-cost microwave integrated circuits (MICs). Among the various printed waveguides that can act as leaky-wave radiators (coplanar guides, slot and strip lines, etc.) [7,20], we can refer to structures derivable from lengths of microstrip. Many different configurations can be employed with microstrips acting as traveling-wave radiators. A first class is based on modulating the dominant mode of the structure with periodic loadings, such as resonant patches or slots (Fig. 6a), and also by varying the lineshape periodically with different meander contours (Fig. 6b) [21]. Even though different solutions have been tested, theory on this topic seems to deserve further studies.

A different way of operation concerns the use of uniform structures acting on higher-order modes that can become leaky for certain values of the parameters involved (Fig. 6c). Analysis of the complex propagation characteristics of the microstrip line shows in fact that, in addition to the dominant quasi-TEM mode, the higher-order modes generally become leaky in suitable frequency ranges [7,20] (see Fig. 2). In particular, it is seen that the

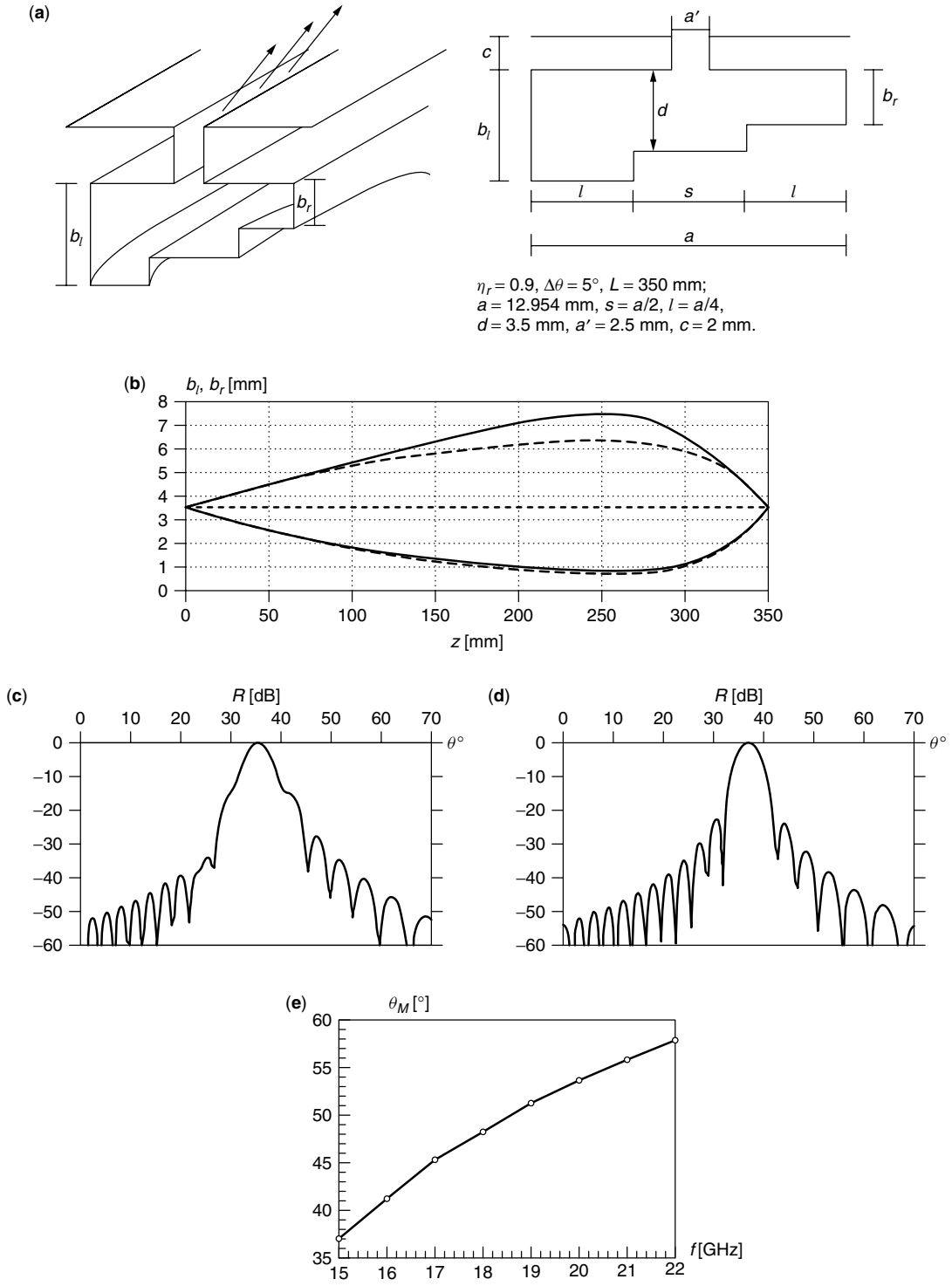


Figure 4. Example of LWAs tapering procedure to achieve a required aperture distribution for pattern shaping: (**a**) reference structure of a stepped rectangular guide LWA with relevant parameters. (**b**) longitudinal modulation of the lateral steps (b_l, b_r vs. z) related to a cosine-type illumination function for a microwave application. The dashed line of b_l, b_r vs. z profile is obtained with a single-shot tapering procedure, that is only an action on the imbalance $\Delta b = (b_l - b_r)/(b_l + b_r)$ taking a constant value of mean height $b_m = (b_l + b_r)/2$ (thus, variations on the phase constant are anyway introduced). The solid-line profile is due to a double-shot tapering procedure, where phase errors are compensated by suitably varying b_m. (**c**) "Distorted" normalized radiation pattern R (dB) according to the dashed-line profile. (**d**) "Correct" radiation pattern according to the solid-line profile for the cosine illumination of the stepped LWA. (**e**) Typical scanning properties for the pointed beam as a function of the frequency (stepped LWA under investigation).

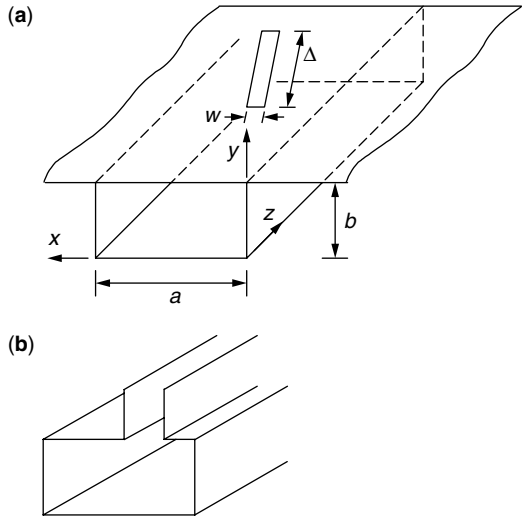


Figure 5. (**a**) Top-wall slitted rectangular guide LWA; (**b**) stub-loaded rectangular guide LWA.

first higher-mode EH_1 can be excited with a proper odd-type source (the mid-plane of symmetry is a perfect electric conductor) and, as frequency is raised, starts to leak power. In general, for the planar structures, leakage can occur in two forms: the surface wave leakage (power that is carried away through the TE and/or TM surface modes of the layered structure), and the "space-wave" leakage (power that is carried away through the TEM mode of the free space) [22]. It is found that, for suitable choices of the parameters with an appropriate excitation, the EH_1 mode can represent rather efficiently the radiation of the microstrip in a certain frequency range (see, e.g., Fig. 2). The coupling phenomenon between the feeding and the radiating line is an aspect to be accurately evaluated, and simplified equivalent networks can be convenient to this aim (23). Radiation performance of printed-circuit LWAs (concerning power handling, polarization, efficiency, pattern shaping, etc.) can be less versatile and satisfactory if compared with LWAs derived from metal guides. From a practical point of view, difficulties can be found particularly in acting independently on the phase and leakage constants through the physical parameters. Uniform-type microstrip LWAs have also been investigated in array configurations for 2D pencil-beam scanning [7,24,25].

4.3. Nonradiative Dielectric (NRD) Guide LWAs

Nonradiative-dielectric (NRD) waveguide, proposed for millimeter-wave applications [26] (Fig. 7a), is a hybrid metal/dielectric guide; it consists of a dielectric rod inserted between metal plates placed at a distance apart that is less than the free-space wavelength. In this way, each discontinuity that preserves the central horizontal-plane symmetry gives only reactive contributions, reducing interference and radiation effects in integrated circuits. A number of passive and active components has been realized with such topology, and also integrated antennas and arrays have been proposed [27,28]. Usually NRD LWAs employ some asymmetry in the basic geometry in order to make leaky the operating mode. A first possible choice [27] (Fig. 7b) is to shorten the length of the plates so that the bound operating mode (LSM_{01}) [26] presents a nonnegligible amplitude contribution on the equivalent aperture, and can give rise to an outgoing leaky wave in the fast-wave range. Another possible choice [7] (Fig. 7c) is to insert some geometrical asymmetry with respect to the central plane (typically an airgap between dielectric and metal), so that a field having a net electric component perpendicular to the plates can be excited, and power can leak out in the form of a TEM-like mode traveling at an angle in the parallel-plate region toward the external environment. Various

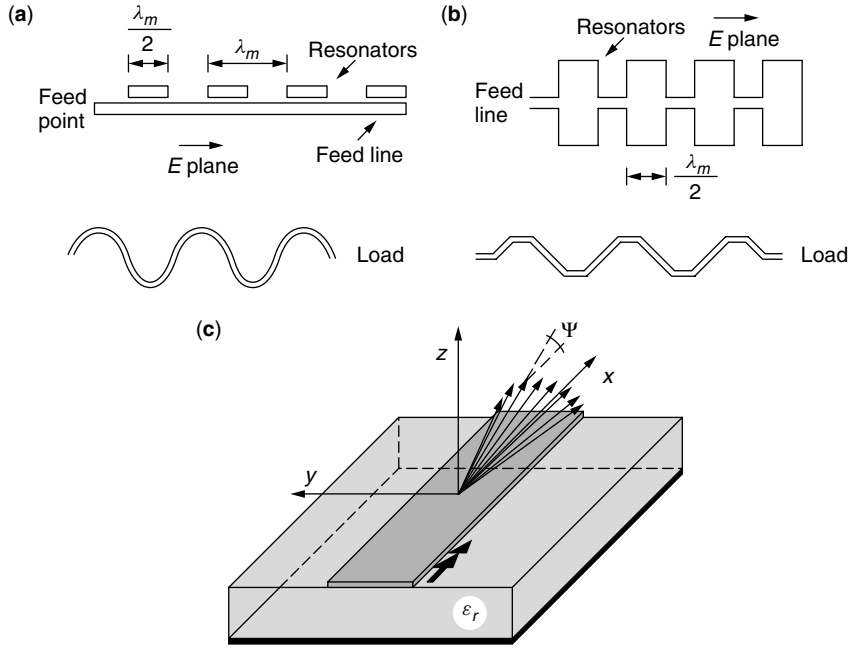


Figure 6. (**a**) Periodically loaded microstrip LWA; (**b**) periodical meander microstrip LWA; (**c**) uniform higher-mode microstrip LWA—space-wave radiation can be associated, for example, with the strip current distribution of the EH_1 mode, which is leaky in a suitable frequency range (see Fig. 2).

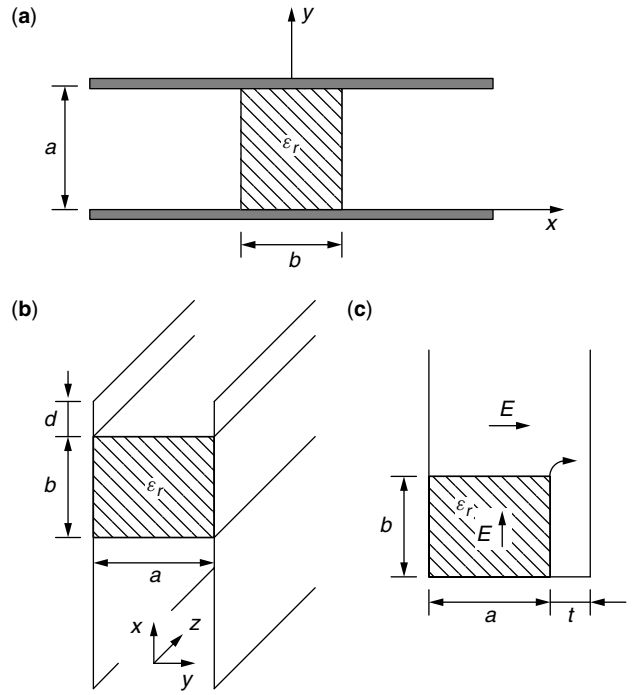


Figure 7. (**a**) Nonradiative dielectric (NRD) waveguide; (**b**) foreshortened NRD LWA; (**c**) asymmetrical NRD LWA.

analyses and design procedures have been developed for these configurations in conjunction with measurements on prototypes.

4.4. Dielectric LWAs

As said, in basic dielectric guides a periodic loading is required in order to isolate a suitable fast-wave space harmonic from the intrinsically slow-wave structure. The most usual periodic perturbation is represented by grating of grooves [29] or metal strips [30,31], usually placed in the top surface of the guide (Fig. 8a); also lateral metal patches can be used in hybrid forms (dielectric/microstrip) (Fig. 8c) [32]. When a sidelobe control is required, the taper is realized on the periodic perturbation (e.g., with grooves or strips slightly changing their dimensions longitudinally). Various studies have been developed to characterize the theoretical performances for these radiators [33]; also, practical aspects have been analyzed, such as the proper feeding elements in order to avoid spurious radiation, and the reduction of the beamwidth in the cross-plane with flared horns [34] (Fig. 8d). All these topologies are good candidates particularly for high-frequency applications (millimeter and submillimeter waves), where the use of dielectric instead of metal for the guidance can reduce the loss effects.

4.5. Layered Dielectric-Guide LWAs

It has been observed that LWAs based on single dielectric layers, also with a ground plane on one side, usually present quite high leakage values, with consequent weak capability in focusing radiation. A significant improvement is achievable by using additional dielectric layers (Fig. 9a); in particular, interesting analyses were performed on substrate/superstrate layered structures [35–37]. By properly dimensioning the heights and the dielectric constants (usually the substrate has lower permittivity than the superstrate), it is possible to excite with a simple element (dipole or slot) a leaky wave giving a conical (due to the symmetries of the topology) highly directive beam. More recently, this basic substrate/superstrate topology has been arranged to allow for a very focused pencil beam with a limited number of radiating elements in form of widely spaced array, exploiting an interaction between leaky and Floquet's modes (Fig. 9b) [38]. Through very simple design procedures, such configurations have the advantages of good radiative performance (high directivity, absence of grating lobes, etc.) with an array

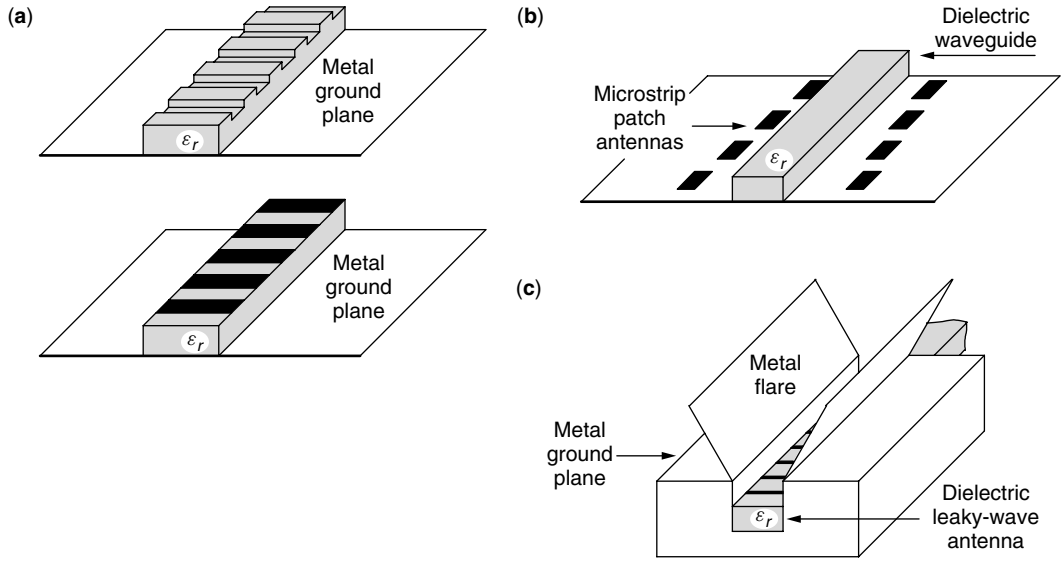


Figure 8. (**a**) Periodically-loaded dielectric LWAs; (**b**) hybrid dielectric/microstrip (insular guide with patches) LWA; (**c**) dielectric LWA with a flared horn to reduce the cross-plane beam width.

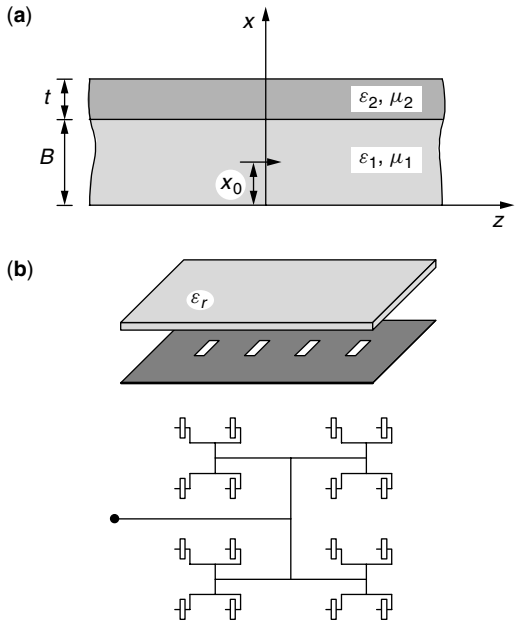


Figure 9. (**a**) Layered dielectric LWA based on a substrate/superstrate structure with a dipole excitation; (**b**) high-gain LW arrays of widely spaced elements in a substrate/superstrate structure: linear and planar configurations (for the latter case, a top view is shown for a microstrip feeding network of widely spaced slot elements on the ground plane of the substrate/superstrate structure).

of few spaced 1D or 2D elements, reducing the cost of the beam forming network and exploiting the greater interspace available at high frequencies (dual-polarization applications, etc.).

5. PRACTICAL CONSIDERATIONS AND MEASUREMENTS

5.1. Feed, Losses, and Manufacture

Feeding LWAs is usually quite simple. In particular, for LWAs derived by metal guides, the feed is represented by a continuous transition from the closed structure acting on a suitable guided mode to the related open one acting on the perturbed (leaky) mode [1–4]. Tapered transitions from the closed to the open structures can be realized to reduce the discontinuity effects and the possible excitation of spurious modes that could arise from abrupt transitions. At the output termination, the introduction of a matched load drastically decreases the remaining power that, if reflected, should give rise to a backlobe, in a direction symmetrical to the main beam with respect to the broadside. The use of dielectric structures can present more difficulties in feeding, in particular in planar configurations. For efficiency and radiation performance, attention has to be paid in avoiding the excitation of additional guided and leaky modes, and also in obtaining a good excitation of the desired leaky wave. For planar guides, such as microstrip or layered dielectrics, local coupling elements (such as slot or dipoles) are usually employed to excite the leaky mode from an input line toward the radiating line.

Ohmic losses do not usually affect much the radiative performance (efficiency, etc.) of LWAs, since the attenuation due to the leakage of radiated power is generally more influent than the attenuation due to dissipated power in the nonideal guiding structure [16]. However, as frequency increases, power loss can be excessive, particularly for LWAs based on closed metal guides. Therefore, for millimeter wave applications the choice of open guides with dielectrics and limited use of metal is often advisable.

The general simplicity of LWA structures makes their manufacture usually easy to perform, even though different construction problems can arise depending on the chosen topology and the frequency range. Simple structures are particularly desirable at millimeter waves, due to the reduced dimensions. On the other hand, too simplified shapes seldom can allow a good control of the radiation performance. In particular, a delicate aspect concerns the usually small longitudinal modifications of the geometry related to tapering for sidelobe control. In this case, an accurate determination of the fabrication imprecisions and tolerances has particular importance to avoid overwhelming the required geometric variations for tapering, thus degrading the improvements of the pattern shaping. Finally, the effects of radomes, used for environmental protection, have also been analyzed [39].

5.2. Measurement Techniques

The radiation properties of LWAs can be tested experimentally through different types of measurement, most of them applicable to aperture antennas [40]. Some basic parameters, such as efficiency and mismatching effects, can be measured directly through the transmission and/or reflection scattering parameters with a network analyzer. Radiation patterns and directivity properties as a function of the observation angles (θ and ϕ in the zenith and azimuth planes, respectively) can be measured for various frequency values with different techniques, on the aperture, in the radiating near field (Fresnel region), and in the far field (Fraunhofer region) [17].

Measurements on the aperture are quite easy to perform, in particular for LWAs derived from partially open metal guides. As already said, the basic parameters to be determined in LWAs, from which a complete knowledge of the radiative characteristics is achieved, are the phase and the leakage constants. A measurement of the field in the close proximity of the aperture can be achieved with a small pickup element (e.g., an electric dipole probe placed parallel to the aperture electric field). Amplitude and phase of the signal received by the probe are thus measurable through a network analyzer, with possible compensations related to the mutual coupling between the current distribution on the aperture and the current probe element.

BIOGRAPHIES

Fabrizio Frezza received the Laurea (degree) cum laude in electronic engineering from "La Sapienza" University of Rome, Italy, in 1986. In 1991 he obtained a doctorate in applied electromagnetics from the same university.

In 1986, he joined the Electronic Engineering Department of the same university, where he has been a researcher from 1990 to 1998, a temporary professor of electromagnetics from 1994 to 1998, and an associate professor since 1998. His main research activity concerns guiding structures, antennas and resonators for microwaves and millimeter waves, numerical methods, scattering, optical propagation, plasma heating, and anisotropic media.

Dr. Frezza is a senior member of IEEE, a member of Sigma Xi, of AEI (Electrical and Electronic Italian Association), of SIOF (Italian Society of Optics and Photonics), of SIMAI (Italian Society for Industrial and Applied Mathematics), and of AIDAA (Italian Society of Aeronautics and Astronautics).

Alessandro Galli received the Laurea degree in electronic engineering in 1990 and a Ph.D. in applied electromagnetics in 1994, both from "La Sapienza" University of Rome, Italy. In 1990, he joined the Electronic Engineering Department of "La Sapienza" University of Rome for his research activity. In 2000, he became temporary professor of electromagnetic fields for telecommunications engineering at "La Sapienza" University of Rome, and in 2002 he became associate professor of electromagnetics at the same university.

His scientific interests mainly involve electromagnetic theory and applications, particularly regarding analysis and design of passive devices and antennas (dielectric and anisotropic waveguides and resonators, leaky-wave antennas, etc.) for microwaves and millimetre waves. He is also active in bioelectromagnetics (modeling of interaction mechanisms with living matter, health safety problems for low-frequency applications, and mobile communications, etc.). In 2000, he was selected as a member of the Technical Committee of the Advisor chosen by the Italian Government for the licenses of the third-generation cellular phones (UMTS).

Dr. Galli is a member of IEEE (the Institute of Electrical and Electronics Engineers). In 1994, he received the Barzilai Prize for the best scientific work of under-35 researchers at the 10th National Meeting of Electromagnetism. In 1994 and 1995, he was the recipient of the Quality Presentation Recognition Award presented by the IEEE Microwave Theory and Techniques Society (MTT-S).

Paolo Lampariello obtained the Laurea degree (cum laude) in electronic engineering at the University of Rome, Italy in 1971.

In 1971, he joined the Institute of Electronics, University of Rome. Since 1976, he has been engaged in educational activities involving electromagnetic field theory. He was made professor of electromagnetic fields in 1986. From November 1988 to October 1994 he served as head of the Department of Electronic Engineering of the "La Sapienza" University of Rome. Since November 1993, he has been the president of the Electronic Engineering Curriculum and since September 1995 he has been the president of the Center Interdepartmental for Scientific Computing. From September 1980 to August 1981 he was a NATO postdoctoral research fellow at the Polytechnic Institute of New York, Brooklyn.

Professor Lampariello has been engaged in research in a wide variety of topics in the microwave field, including electromagnetic and elastic wave propagation in anisotropic media, thermal effects of electromagnetic waves, network representations of microwave structures, guided-wave theory with stress on surface waves and leaky waves, traveling-wave antennas, phased arrays, and, more recently, guiding and radiating structures for the millimeter and near-millimeter wave ranges.

Professor Lampariello is a fellow of the Institute of Electrical and Electronics Engineers, and a member of the Associazione Elettrotecnica ed Elettronica Italiana.

BIBLIOGRAPHY

1. A. A. Oliner, Leaky-wave antennas, in R. C. Johnson, ed., *Antenna Engineering Handbook*, 3rd ed., McGraw-Hill, New York, 1993, Chap. 10.
2. C. H. Walter, *Traveling Wave Antennas*, McGraw-Hill, New York, 1965; Peninsula Publishing, Los Altos, CA, reprint, 1990.
3. T. Tamir and A. A. Oliner, Guided complex waves, Parts I and II, *Proc. IEE* **110**: 310–334 (1963).
4. T. Tamir, Inhomogeneous wave types at planar interfaces: III — Leaky waves, *Optik* **38**: 269–297 (1973).
5. I. Ohtera, Diverging/focusing of electromagnetic waves by utilizing the curved leakywave structure: Application to broad-beam antenna for radiating within specified wide-angle, *IEEE Trans. Antennas Propag.* **AP-47**: 1470–1475 (1999).
6. R. C. Hansen, *Phased Array Antennas*, Wiley, New York, 1998.
7. A. A. Oliner (principal investigator), *Scannable Millimeter Wave Arrays*, Final Report on RADC Contract F19628-84-K-0025, Polytechnic Univ., New York, 1988.
8. T. Itoh, Millimeter-wave leaky-wave antennas, *Proc. Int. Workshop Millimeter Waves*, Italy, 1996, pp. 58–78.
9. T. Itoh, ed., *Numerical Techniques for Microwave and Millimeter-Wave Passive Structures*, Chap. 3 (J. R. Mosig), Chap. 5 (T. Uwano and T. Itoh), and Chap. 11 (R. Sorrentino), Wiley, New York, 1989.
10. R. Sorrentino, ed., *Numerical Methods for Passive Microwave and Millimeter Wave Structures*, IEEE Press, New York, 1989.
11. L. O. Goldstone and A. A. Oliner, Leaky-wave antennas — Part I: Rectangular waveguides, *IRE Trans. Antennas Propag.* **AP-7**: 307–319 (1959).
12. N. Marcuvitz, *Waveguide Handbook*, McGraw-Hill, New York, 1951.
13. P. Lampariello, F. Frezza, and A. A. Oliner, The transition region between bound-wave and leaky-wave ranges for a partially dielectric-loaded open guiding structure, *IEEE Trans. Microwave Theory Tech.* **MTT-38**: 1831–1836 (1990).
14. S. Majumder, D. R. Jackson, A. A. Oliner, and M. Guglielmi, The nature of the spectral gap for leaky waves on a periodic strip-grating structure, *IEEE Trans. Microwave Theory Tech.* **MTT-45**: 2296–2307 (1997).

15. R. E. Collin, *Field Theory of Guided Waves*, 2nd ed., IEEE Press, New York, 1991.

16. C. Di Nallo, F. Frezza, A. Galli, and P. Lampariello, Rigorous evaluation of ohmic-loss effects for accurate design of traveling-wave antennas, *J. Electromagn. Wave Appl.* **12**: 39–58 (1998).

17. C. Di Nallo et al., Stepped leaky-wave antennas for microwave and millimeter-wave applications, *Ann. Télécommun.* **52**: 202–208 (1997).

18. F. L. Whetten and C. A. Balanis, Meandering long slot leaky-wave waveguide antennas, *IEEE Trans. Antennas Propag.* **AP-39**: 1553–1560 (1991).

19. P. Lampariello et al., A versatile leaky-wave antenna based on stub-loaded rectangular waveguide: Parts I–III, *IEEE Trans. Antennas Propag.* **AP-46**: 1032–1055 (1998).

20. H. Shigesawa, M. Tsuji, and A. A. Oliner, New improper real and complex solutions for printed-circuit transmission lines and their influence on physical effects, *Radio Sci.* **31**: 1639–1649 (1996).

21. J. R. James and P. S. Hall, *Handbook of Microstrip Antennas*, Peter Peregrinus, London, 1989.

22. F. Mesa, C. Di Nallo, and D. R. Jackson, The theory of surface-wave and space-wave leaky-mode excitation on microstrip lines, *IEEE Trans. Microwave Theory Tech.* **MTT-47**: 207–215 (1999).

23. P. Burghignoli et al., An unconventional circuit model for an efficient description of impedance and radiation features in printed-circuit leaky-wave structures, *IEEE Trans. Microwave Theory Tech.* **MTT-48**: 1661–1672 (2000).

24. C. N. Hu and C. K. C. Tzuang, Microstrip leaky-mode antenna array, *IEEE Trans. Antennas Propag.* **AP-45**: 1698–1699 (1997).

25. P. Baccarelli et al., Full-wave analysis of printed leaky-wave phased arrays, *Int. J. RF Microwave Comput. Aid. Eng.* (in press).

26. T. Yoneyama, Nonradiative dielectric waveguide, in K. J. Button, ed., *Infrared and Millimeter-Waves*, Academic Press, New York, 1984, Vol. 11, pp. 61–98.

27. A. Sanchez and A. A. Oliner, A new leaky waveguide for millimeter waves using nonradiative dielectric (NRD) waveguide — Parts I and II, *IEEE Trans. Microwave Theory Tech.* **MTT-35**: 737–752 (1987).

28. J. A. G. Malherbe, An array of coupled nonradiative dielectric waveguide radiators, *IEEE Trans. Antennas Propag.* **AP-46**: 1121–1125 (1998).

29. F. Schwering and S. T. Peng, Design of dielectric grating antennas for millimeter-wave applications, *IEEE Trans. Microwave Theory Tech.* **MTT-31**: 199–209 (1983).

30. M. Ghomi, B. Lejay, J. L. Amalric, and H. Baudrand, Radiation characteristics of uniform and nonuniform dielectric leaky-wave antennas, *IEEE Trans. Antennas Propag.* **AP-41**: 1177–1186 (1998).

31. S. Kobayashi, R. Lampe, R. Mittra, and S. Ray, Dielectric-rod leaky-wave antennas for millimeter-wave applications, *IEEE Trans. Antennas Propag.* **AP-29**: 822–824 (1981).

32. A. Henderson, A. E. England, and J. R. James, New low-loss millimeter-wave hybrid microstrip antenna array, *Proc. 11th Eur. Microwave Conf.*, 1981, pp. 825–830.

33. M. Guglielmi and A. A. Oliner, Multimode network description of a planar periodic metal-strip grating at a dielectric interface — Parts I and II, *IEEE Trans. Microwave Theory Tech.* **MTT-37**: 534–552 (1989).

34. T. N. Trinh, R. Mittra, and R. J. Paleta, Horn image-guide leaky-wave antenna, *IEEE Trans. Microwave Theory Tech.* **MTT-29**: 1310–1314 (1981).

35. D. R. Jackson and N. G. Alexopoulos, Gain enhancement methods for printed circuit antennas, *IEEE Trans. Antennas Propag.* **AP-33**: 976–987 (1985).

36. D. R. Jackson and A. A. Oliner, A leaky-wave analysis of the high-gain printed antenna configuration, *IEEE Trans. Antennas Propag.* **AP-36**: 905–910 (1988).

37. H. Ostner, J. Detlefsen, and D. R. Jackson, Radiation from one-dimensional dielectric leaky-wave antennas, *IEEE Trans. Antennas Propag.* **AP-43**: 331–339 (1995).

38. L. Borselli, C. Di Nallo, A. Galli, and S. Maci, Arrays with widely-spaced high-gain planar elements, *1998 IEEE AP-S Int. Symp. Dig.*, 1998, pp. 1446–1449.

39. C. Di Nallo, F. Frezza, A. Galli, and P. Lampariello, Analysis of the propagation and leakage effects for various classes of traveling-wave sources in the presence of covering dielectric layers, *1997 IEEE MTT-S Int. Microwave Symp. Dig.*, 1997, pp. 605–608.

40. C. A. Balanis, *Antenna Theory: Analysis and Design*, Wiley, New York, 1997, Chap. 16.

LEO SATELLITE NETWORKS

THOMAS R. HENDERSON
Boeing Phantom Works
Seattle, Washington

1. INTRODUCTION

Since the mid-1960s, most communications satellites have been deployed in a geostationary orbit, so named because the satellite appears to an earth-bound observer to remain nearly fixed in the sky. The geostationary orbit is a circular equatorial orbit at an altitude of 35,786 km, in which the angular velocity and direction of the satellite matches the angular rate of the rotation of the earth's surface. Satellites in this orbit provide telecommunications trunking services, VSAT (very-small-aperture terminal) data networks, direct-to-home television broadcasts, and even mobile services.

Although the very first satellites were launched into low orbits (since lower orbits were cheaper and less risky to attain), the convenience of geostationary orbits soon became the dominant factor in orbit selection. However, the latter half of the 1990s witnessed a renewal of interest in deploying communications satellites in orbits much closer to the earth [hence the term *low-earth-orbit* (LEO)], driven by the desire to extend voice, low-speed data, and Internet access services to mobile or remote users. Satellites at lower orbits have the drawback that they do not appear fixed in the sky. To provide continuous coverage within a given service region, more than one satellite (i.e., a network or *constellation* of satellites) is needed. Several such commercial systems have already been deployed (most notably the Iridium [1] and Globalstar [2] systems),

and even more ambitious systems have been proposed. Satellite constellations at lower orbits offer the following primary advantages:

- The end-to-end propagation delays can be significantly lower, thereby improving the quality of service provided to voice-based and data applications.
- Advances in cellular telephony electronics for handheld devices have enabled truly handheld satellite terminals equipped with small low-gain antennas, reachable by satellites at these lower orbits.
- As the orbital altitude increases, it is necessary to use larger antennas onboard to support the small spot beam (i.e., cell) sizes required for large system capacities.

This article summarizes the key issues regarding satellite networks employing LEO or other nongeostationary orbits. We first describe the various orbital geometries available to the satellite system designer. Next, we highlight several differences between satellite systems designed using geostationary (GSO or GEO) and nongeostationary orbits (non-GSO). Finally, we describe networking issues that arise from the need to use many satellites over time to serve terminals on the ground.

Our focus is on satellite networks that provide *continuous communications services* to a given service region. Therefore, we will not be explicitly focusing on store-and-forward satellite communications networks, or on satellites used for remote sensing, position determination, or military purposes, although many of the same principles apply.

2. SATELLITE ORBITS

2.1. Basic Orbital Geometry

To first order, a satellite's orbit can be described by an ellipse lying in a fixed orbital plane, with the earth's center positioned at one of the foci of this ellipse. As the satellite proceeds around this orbit, the earth rotates underneath it. The combination of the earth's rotation and the satellite's movement within the orbital plane contribute to its apparent motion in the sky as viewed from earth. The shape of the orbit is defined by its eccentricity (e) and its semimajor axis (a). The point at which the satellite is furthest from the earth is known as the *apogee* and, conversely, the closest point is the *perigee*. Additionally, there are three parameters that describe the orientation of this ellipse with respect to the earth. The right ascension of the ascending node (Ω) is a positive angle measured in the equatorial plane between two directions — a reference direction in the coordinate system, and the direction of the ascending node. The reference direction is given by intersection of the equatorial plane and the plane of the ecliptic, and is known as the direction of vernal equinox.[1]

[1] This reference direction maintains a fixed orientation in space with time and is so named because passes through both the earth and the sun on the vernal (spring) equinox.

The *ascending node* is the point of intersection between the orbital plane and the plane of the equator, the satellite crossing this plane from south to north. The inclination (i) is the positive angle between the normal to the direction of ascending node (pointed toward the east) in the equatorial plane and the normal to the line of nodes (in the direction of the velocity) in the orbital plane. The inclination can range from 0° to 180°; orbits with inclination greater than 90° are called *retrograde* orbits. The argument of perigee (ω) defines how the elliptical orbit is oriented in the plane. It is defined as the positive angle in the orbital plane between the direction of ascending node and direction of perigee (w ranges from 0° to 360°). A sixth parameter, the time of perigee passage (τ), defines the position of the satellite within this orbit (i.e., it specifies an initial condition). The period of the orbit (T) is given by the following relationship

$$T = 2\pi \left(\frac{a^3}{\mu}\right)^{1/2} \quad \text{(s)} \tag{1}$$

where $\mu = 3.9866 \times 10^{14}$ m^3/s^2 is the gravitational parameter for the earth, and a is the semimajor axis. Figure 1 illustrates these orbital parameters.

There are a variety of orbital perturbations (asymmetry of terrestrial gravitational potential due to the earth's oblateness, solar radiation pressure, solar and lunar gravitational influences, and atmospheric drag) that cause the actual orbit to deviate from this idealized model. To counteract such perturbations, satellites periodically apply controlled motor thrusts in a process known as *station keeping*. Station keeping requires the storage onboard of excess fuel reserves (such as pressurized nitrogen), the quantity of which may determine the operating lifetime of the satellite since they are not replenishable. To ensure that the satellite constellation geometry can remain fixed in the face of such

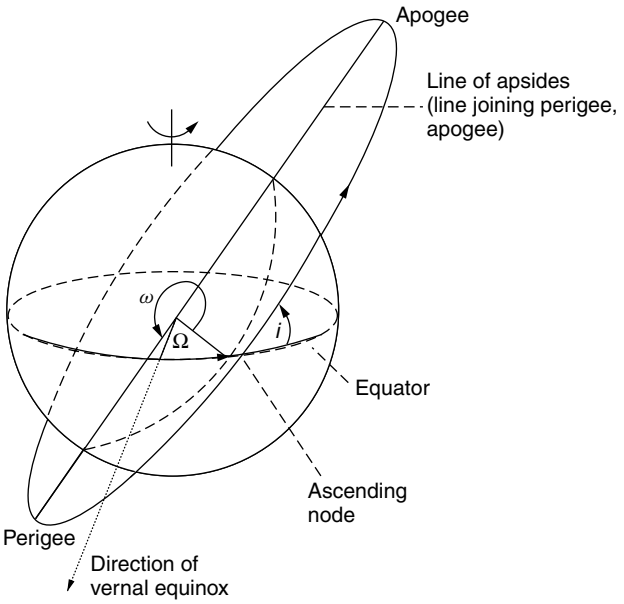


Figure 1. Illustration of Keplerian orbital parameters that define a satellite orbit (from Pattan [3], used with permission).

perturbations, all satellites in a given constellation should assume the same inclination and altitude [3]. However, even under these preferred geometries, the network operator may choose not to expend the fuel required to maintain the relative distances between satellites in the same orbital plane or in different orbits. Constellations that maintain these relationships are known as *phased* constellations, and provide more optimal solutions to the network design, at the expense of requiring larger satellites and a control station network [4]. The book by Maral and Bousquet [5] provides an extensive treatment of orbital perturbations.

2.2. Types of Orbit

In principle, satellites may be deployed into any orbit, but there are practical and economic considerations that favor certain orbit types over others. The choice of orbital configuration and number of satellites is the result of a system optimization combining a large number of factors that we will summarize shortly. As far as the orbits themselves, the International Telecommunications Union (ITU) has defined three broad classifications for nongeostationary orbits:

- *Low-Earth Orbit* (*LEO*). LEO satellite orbits lie between roughly 700 and 1500 km. The lower altitude bound is governed by limits on atmospheric drag, while the upper bound is the approximate beginning of the inner Van Allen radiation belt.[2] LEO orbits are typically circular. For coverage of the entire earth's surface, a near-polar inclination can be selected — this, however, causes a concentration of coverage near the poles. If the inclination angle is relaxed, a higher degree of system capacity can be concentrated at midlatitudes, at the expense of polar coverage.
- *Medium-Earth Orbit* (*MEO*). Systems that lie between the two Van Allen radiation belts (between 5000 and 13,000 km), or above the outer belt (greater than 20,000 km) are typically classified as MEO satellite systems. The term *intermediate circular orbit* (ICO) is also sometimes used when the orbit is circular. Because of their use of a higher altitude, MEO systems require fewer satellites than do LEO systems to provide similar coverage. For example, the Iridium (LEO) system uses 66 active satellites for global coverage, while the commercially proposed ICO constellation [6] requires only 10.
- *Highly Elliptical Orbit* (*HEO*). A third option has been to use elliptical, inclined orbits. The key property of an elliptical orbit is that the velocity of the orbit is not constant but instead is slowest at the orbit apogee.[3] Therefore, satellites in such orbits can remain visible for longer stretches of time if the apogee is situated over the desired region of coverage. Furthermore, unlike GSO satellites, HEO satellites can serve latitudes higher than 75°. One drawback to elliptical orbits is that the oblateness of the earth, and the resulting anomalies in the gravitational field, causes the apogee to rotate slowly around the orbit (a phenomenon known as *apsidal rotation*). There are, however, two orbital inclinations (63.4° and 116.6°) for which no apsidal rotation occurs. One such orbit, known as the *Molnya* (lightning) orbit, uses an inclination of 63.4°. Molnya orbits, pioneered by the former Soviet Union, have an apogee at roughly 40,000 km, a perigee of about 1000 km, an argument of perigee of about 270°, and a period of 12 h. By using multiple satellites in such orbits and ground stations that can track the slowly moving satellites, communications at high latitudes can be enabled with a high elevation angle over the horizon. Note that these orbits must pass through the Van Allen radiation belts. Typically, this requires more radiation shielding of the electronics and results in a shorter satellite lifetime; as a result, variations to this orbit that do not require crossing the radiation belts have been studied.

2.3. Coverage

The maximal satellite *footprint*, or coverage area, is governed by the altitude above the earth's surface and the minimum elevation angle supported. Details on the geometry of this relationship are covered by Maral and Bousquet [5]. The actual coverage area may be smaller if the antenna pattern is more focused on a smaller surface area. Furthermore, the coverage area is usually segmented into a collection of smaller *spot beams*. This is done primarily for two reasons: (1) as in cellular networks, the overall system capacity can be increased through frequency reuse — for example, in the Iridium system, the satellite footprint is divided into 48 smaller spot beams, with a frequency reuse factor of 12 [1]; and (2) the communications link performance is inversely related to the spot size illuminated, because smaller spot beams result in more focused RF carrier power. The costs of supporting smaller spot beams include larger aperture antennas on board the satellite, more frequent link handoffs for terminals, and a more sophisticated payload to route traffic to the correct spot beam if onboard switching is performed.

2.4. Constellation Design

Satellite constellations are typically designed based on a requirement of having one or more satellites continuously in view of earth stations (above some minimum elevation angle) throughout a given service area. One of the main objectives is to minimize the number of satellites needed

[2] The Van Allen radiation belts consist of two toroidally shaped regions around the earth's magnetic equator where highly charged particles are trapped by the magnetic field. The inner belt lies between approximately 1500 and 5000 km, and the outer belt between 13,000 and 20,000 km. It is preferable to avoid prolonged exposure to such regions because of damaging effects on solid-state electronics.

[3] This is a consequence of Kepler's second law of planetary motion, which states that the radius vector of the orbit sweeps out equal areas in equal times (the "law of areas").

to meet this requirement. Walker originally explored different types of constellations using circular orbits [7], which are generally classified into two categories. The first category, constellations with orbits using near-polar inclination (sometimes called *Walker star* or *polar* constellations), have the property that the ascending nodes of the orbits are regularly distributed over a hemisphere (180°). As a result, there are two hemispheres in which all the satellite orbits are corotating, in either a north–south or south–north direction. The division between these hemispheres of coverage, across which the satellite orbits are counterrotating, is commonly called a *seam*. Although this type of constellation has a concentration of coverage at the poles, it efficiently covers the lower latitudes, and has the desirable property that satellites in corotating planes move slowly with respect to one another, allowing for easier establishment of intersatellite communications links between them. The Iridium constellation uses a design of this type, as illustrated in Fig. 2.

The second category of circular-orbiting constellations, known as *Walker delta* or *rosette* constellations, have the ascending nodes distributed uniformly across 360° of longitude, with the orbits all at the same inclination. The result of this design is that any area of the earth's surface has both ascending and descending satellites. This type of constellation design is most commonly applied when the inclination angle is relaxed below 90°, so that coverage can be concentrated at the populated midlatitudes. When the satellites are connected via intersatellite links, this constellation design also offers networking path diversity not achievable with polar constellations [9]. Globalstar uses a rosette constellation design, as illustrated in Fig. 3. The book by Pattan provides further details on polar and rosette constellation designs [3].

As emphasized above, nongeostationary satellites are not limited to circular orbits. Draim has derived optimal constellation geometries based on elliptical orbits, the principles of which have been incorporated into the proposed Ellipso constellation [10]. We have already introduced the Molnya orbit as an example of a highly elliptical orbit. Another elliptical orbit known as the *Tundra* orbit shares the same orbital inclination of 63.4° but with a period of 24 h. Since the visibility of each Tundra satellite beneath the apogee is greater than 12 h, two satellites in such orbits are sufficient for continuous coverage [5].

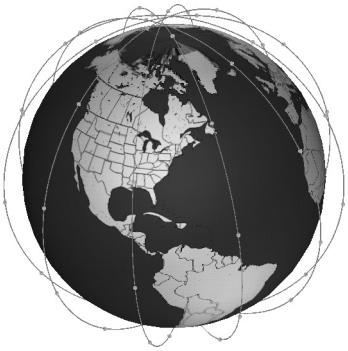


Figure 2. Orbital geometry for the Iridium constellation, an example of a polar-orbiting constellation (from Thurman and Worfolk [8], used with permission).

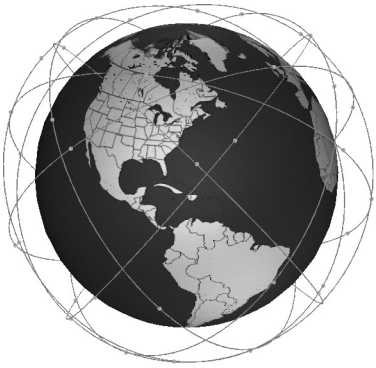


Figure 3. Orbital geometry for the Globalstar constellation, an example of a rosette constellation (from Thurman and Worfolk [8], used with permission).

3. GEOSTATIONARY VERSUS NONGEOSTATIONARY SATELLITES

In this section, we highlight some of the distinguishing properties of nongeostationary satellites by contrasting them with geostationary satellites.

3.1. Propagation Delay

One of the reasons most frequently cited for moving to lower-earth orbits is the resultant reduction in propagation delay. For GSO satellites, the one-way propagation delay from the earth to the satellite is between 120 and 140 ms. This makes the round-trip delay based on propagation delay alone somewhere between 480 and 560 ms, and additional delays for coding, queueing, and multiple access typically push this number above 600 ms. For voice traffic, this amount of delay tends to disrupt the rhythms of conversation and can lead to annoying echos when analog phones are involved. Furthermore, delays of this magnitude can seriously compromise the throughput performance of Internet transport protocols, as well as affect interactive data applications. However, for other applications such as broadcasting, the delay is not important. Similar delays are experienced when using satellites in the Molnya orbit.

By moving satellites to lower orbits, the resultant round-trip delays can be as low as 10–20 ms — similar to what is experienced over the wide-area Internet. However, it must be emphasized that this is only a lower bound; the actual delay is a function of distance between terminals and the number of satellites traversed and can vary over time as the relative positions change. Furthermore, at low link rates, the delays to access the channel (multiple access) can be a significant component. Analysis of the Iridium system has shown that the average end-to-end delay encountered is roughly 100–200 ms, with a large portion of this delay due to multiple access [1]. Moreover, while the delay is constantly varying due to orbital motions, it also is subject to step changes as link handoffs cause a reconfiguration of the communications

path. Nevertheless, as long as end-to-end delays can be kept in the region of 100 ms or less, the qualitative performance improvement for both voice and data traffic can be substantial.

3.2. Link Performance Issues

The error performance of a radio link is a function of the carrier power transferred (by the antenna) to the receiver input, interference from undesired signals also captured by the antenna, and noise power within the receiver. Many books on satellite communications provide an extensive treatment of link budget issues [e.g., 5]. Many of the same principles apply to nongeostationary channels, but there are some significant differences. Geostationary satellite links providing service to fixed terminals can generally be modeled as additive white Gaussian noise (AWGN) channels, with a fixed free-space path loss. In contrast, LEO channels are subject to Rician fading (particularly if low-gain antennas are used by the terminal), high Doppler shifts, variable path loss (unless compensated for by the spacecraft), and irregular terrestrial interference [11,12]. A popular model for the LEO channel, combining Rice and lognormal statistics, is due to Corazza [13]. Because of the severe fading, researchers have shown the benefit of satellite diversity in improving overall system availability and capacity [14].

3.3. Interference

Geostationary satellites using the same frequency bands and carrier polarizations are typically spaced about two to four degrees apart in the orbital arc. For fixed satellite service, this dictates a minimum size of the terrestrial antenna such that the desired pattern directivity can be maintained. Satellite systems can therefore share frequency bands by exploiting these fixed geometries and directional antennas. The situation becomes considerably more complex when satellites appear to move in the sky, and when mobile handsets (with low directivity antennas) are being used. Satellites and terminals from LEO and GSO systems using the same frequencies are likely to interfere with one another as their relative geometries change. While this can be alleviated by placing LEO and GSO systems at different frequency bands, LEO systems attempting to share the same frequency bands are still likely to interfere with one another. Two possible solutions to this problem are to employ spread-spectrum modulation techniques using code sets with low cross-correlation, or to simply divide the available spectrum among users and have each operate an independent system. For systems with user terminals employing low-gain antennas, the current consensus seems to be that spectrum separation is required, while the jury is still out for broadband systems using terminals with highly directive antennas. In summary, the international regulatory procedures required to operate a non-GSO system are considerable.

3.4. Frequencies

Because satellite orbits and frequencies do not belong to any nationality exclusively, their use is coordinated by the ITU. It should be emphasized that the particular allocations change over time and the details are complicated, so we will simply provide an overview of the main frequency bands herein. Briefly, the ITU has established that geostationary satellite links use frequencies found mainly in the L (roughly 1.5 GHz downlink, 1.6 GHz uplink), C (4/6 GHz), Ku (12/14 GHz), and Ka (20/30 GHz) bands. The ITU further classifies satellite systems as providing either fixed satellite service (FSS) (to fixed terminals on the ground), broadcast satellite service (BSS), or mobile satellite service (MSS). Frequency allocations for nongeostationary FSS systems are found in the same general frequency bands used by geostationary satellites. For mobile satellite service, links carrying subscriber traffic can be categorized as either *feeder links* or *subscriber links*. Feeder links connect the satellites to a gateway earth station; these types of links also have been allocated frequencies in the C, Ku, and Ka bands. However, because of the low-gain antennas typical of mobile handsets, there is a strong incentive to use as low a frequency as possible for link performance issues. Specifically, roughly 4 MHz of spectrum in the VHF/UHF bands (around 150 and 400 MHz) and 32 MHz of spectrum in the L band (1.6/2.5 GHz) are allocated to nongeostationary MSS systems.

In the United States, the spectrum in the VHF/UHF bands has been set aside for low-data-rate data systems. Such systems have been coined "little LEOs"; an example is the Orbcomm system used for paging and short data messaging. "Big LEO" systems such as Iridium and Globalstar use the L-band frequencies and are permitted to offer both voice and data services. LEO systems offering broadband data rates will use frequencies in the Ku and Ka bands, or even higher frequencies. The book by Pattan [3] discusses the various frequency allocations in more detail.

3.5. Launch and Spacecraft

As the orbital altitude is increased, the cost of deploying a satellite into that orbit also increases. The geostationary orbit is expensive to attain, requiring a multistep approach. The first step involves placing the satellite into a circular low-earth orbit, then into an elliptical geostationary transfer orbit (where the apogee of this orbit corresponds to the altitude of the geostationary orbit), then into its final orbit. Since only a small fraction of the mass deployed at low-earth orbit is eventually deployed in the final orbit (the rest is fuel), the cost penalty to achieve geostationary orbit is substantial. In contrast, LEO satellites can be launched directly to their final orbital altitude, and for small satellites, multiple satellites can often be launched using the same vehicle. However, while the cost of launching an individual LEO satellite is cheaper than a GSO satellite, the cost of launching a whole constellation is seldom. At least one launch is typically required for each orbital plane of the constellation.

A detailed treatment in the difference between geostationary and nongeostationary spacecraft is beyond this article's scope; the interested reader is directed to the overview in Ref. 5. However, we note that since LEO satellites are closer to the earth, they more frequently undergo shadowing by the earth, and therefore

are subjected to frequent thermal stresses and require batteries to continue to operate while within the shadow. LEO satellites can also be smaller in some dimension (antenna size or transmit power) than GSO satellites while still providing an equivalent RF carrier flux density on the ground. However, the perception that LEO satellites are much smaller than GSO satellites is not necessarily valid in general, but rather is due to the initially deployed LEO systems using a small amount of spectrum. The size of a satellite is directly related to the power required, which is directly related to the throughput; consequently, the proposed broadband LEO satellites plan to use very large satellites.

3.6. Tracking and Link Handoff

Handoff (also known as *handover*) is defined as the procedure for changing the radio communications path to maintain an active communications session. The most significant challenge for a nongeostationary satellite system is the need to track satellites and to hand off active communications links from one satellite to another or between different beams of the same satellite. The rate at which a ground terminal must hand off a connection between satellites (*intersatellite* handoff) varies with the altitude, ranging from 12 h (HEO orbits of type Tundra) to 10 min (LEO). However, handoffs can be even more frequent if the coverage area of the satellite is further segmented into spot beams. For example, the Iridium satellites employ 48 spot beams within the coverage area of one satellite. In this case, handoff between beams on the same satellite can occur every minute or two.

Beam handoffs typically require a change in carrier frequency (unless spread-spectrum modulation is used) and acquisition of the new link. Intersatellite handoffs may require the additional step of repointing the terminal's antenna, which could cause an interruption of service. Such an interruption may be avoided in one of several ways. One brute-force method is to equip the terminal with two mechanically or electronically steered antennas, and engage the nonactive antenna in finding the next satellite. Depending on the service, if a single electronically steered beam is used, the switchover may occur rapidly enough, especially if the approximate position of the next satellite is known by the terminal.

Satellite antenna patterns are typically nadir-pointing, which means that the pattern drags across the surface of the earth with a constant velocity. As a result, handoffs are asynchronous in the system — there will always be some subset of user terminals in the process of handing off at any given time. A proposed alternative would be to electronically or mechanically steer the antenna on board the satellite to keep the coverage fixed on the earth's surface until some point in time at which all of the patterns synchronously switch. This proposed technique would reduce or eliminate intrasatellite handoffs and would cause all intersatellite handoffs to occur synchronously, thereby simplifying the algorithms that deal with handoff [15].

3.7. Intersatellite Links

At low orbital altitudes, the satellite footprint may be relatively small. In a communications session, if both ground terminals are not within the same footprint, some means of transmitting signals between the satellites is necessary. If there are gateway stations located in each satellite's footprint, then one solution is to route traffic from an earth station to a gateway in each footprint, and then to use landlines to interconnect the gateways. Such an approach, while greatly simplifying the satellite payload, has the drawback of requiring a large network of ground-based gateway stations interconnected by terrestrial links.[4]

An alternative solution is to use communications links to interconnect the satellites themselves. These links, known as *intersatellite links* (ISLs), create a mesh network in the sky, and obviate the need to have gateway stations in every coverage footprint (note that this advantage diminishes for MEO/ICO satellites, which have broader coverage footprints). Each Iridium satellite, for example, has ISLs to the two closest satellites within the same orbital plane (black lines illustrated in Fig. 4), and either one or two links to the nearest neighboring satellite in an adjacent plane (lighter lines in Fig. 4). The drawbacks to using ISLs are an increase in complexity of the satellite payload, the establishment and maintenance of such links, as well as the requirement to route traffic between satellites.

The frequencies allocated for ISLs correspond to strong absorption by the atmosphere (to protect against terrestrial interference). Selected radio links at frequencies between 23 and 58 GHz and optical wavelengths between 0.8 and 10.6 μm may be used. For high-capacity links, optical link hardware requires less mass and power consumption.

ISLs require steerable antennas for link pointing, acquisition, and tracking. ISLs between satellites in the same orbiting plane (known as *intraplane* ISLs) do not require tracking in a phased constellation, because the orbital relationship between such satellites is fixed. ISLs that connect satellites in different orbital planes (*interplane* ISLs) will require tracking. The pointing requirements depend strongly on the constellation design. As an example, the Iridium constellation requires a pointing range of roughly 10° in the vertical direction and 140° in the horizontal direction [16]. Furthermore, the pointing angles may become so severe that ISLs will need to be deactivated for a portion of an orbit, or handed off to another satellite. This condition holds in the high-latitude regions of polar-orbiting constellations. Finally, ISLs may be handed off from one satellite to another if the relative locations of the satellites change with respect to one another; for example, ISLs connecting satellites across the seam of a polar-orbiting constellation. As we discuss in the next section, such ISL link changes have implications on network routing.

4. NETWORKING CONSIDERATIONS FOR SATELLITE CONSTELLATIONS

Satellite constellations are considerably more complicated than geostationary satellites from a networking

[4] This approach is used by the Globalstar system.

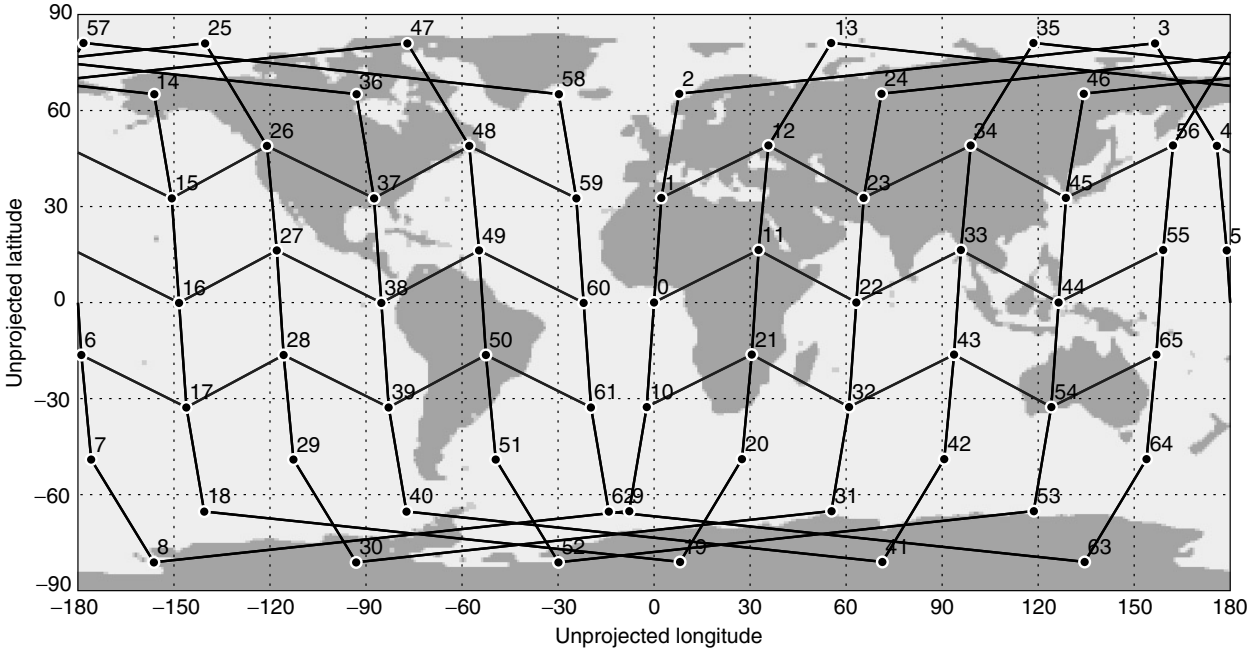


Figure 4. Snapshot of Iridium satellites and active ISLs on an unprojected map of the earth's surface. Lighter lines indicate interplane ISLs; black lines denote intraplane ISLs.

standpoint. Geostationary satellites were originally used as repeaters ("bent pipes") for fixed or semipermanent communications channels. With the desire to share bandwidth among many users, multiple-access protocols such as time-division multiple access (TDMA) and ALOHA[5] were developed. VSAT networks, for example, rely on the coordinated sharing of uplink capacity among hundreds or thousands of terminals. More recently, the construction of multibeam satellites has led to onboard switching between transponders. Only relatively recently have satellite payloads with demodulation, baseband processing, and signal regeneration begun to be deployed. Nevertheless, even with sophisticated onboard processing, geostationary satellites are still not much more than a (fixed location) switch in the sky.

In contrast, satellite constellations are an interesting variant of mobile networks — one in which the network nodes move and the terminals stay (relatively) fixed. There is a mixture of permanent (intraplane ISL), semipermanent (interplane ISL), and transient (ground-to-satellite) links. Unlike traditional mobile networks, however, the network topology is somewhat regular in structure, and many topological changes can be predicted in advance. The regularity and predictability of the network geometry can be exploited in the network architecture design.

In this section, we survey the different approaches that have been proposed for networking in satellite constellations. Since satellite networks are designed to extend services of terrestrial networks, it should not be surprising that satellite network architectures can generally be classified as either circuit switching or packet switching. Before examining both of these approaches in turn, we first note that some common architectural principles apply to both types of networks. The first principle is *flexibility*. Satellite networks are expected to last many years and are difficult or prohibitively expensive to upgrade in space. Therefore, system designers strive to implement general solutions for the space segment that will not become obsolete. For example, packet-switched satellite architectures will likely not implement pure IP (Internet Protocol) packet switching in space, but instead will strive for a generic, satellite-optimized packet switching infrastructure into which IP and other protocols can be mapped. A second related principle is *simplicity*, which argues for deploying functions, when possible, within the ground segment so as to relieve the onboard complexity of the electronics, and hence the mass and power requirements of the spacecraft.

4.1. Circuit-Switched Architectures

Many LEO and MEO satellite systems have been positioned to offer traditional PSTN services (voice, low-bit-rate data, facsimile, paging, etc.). The architectures for these networks typically resemble those of "second generation" mobile communications systems such as GSM. The functions required of the network include basic routing and switching of the call, mobility management, privacy, security, and radio resource management. The chief difference between satellite constellations and traditional mobile networks lies in the mobility management function. As mentioned above, the network nodes, as well as the terminals, in fact move. This has implications on location registration and handoff [17].

Location registration is the procedure by which mobile units are identified as being in a particular location so

[5] ALOHA was developed by N. Abramson at the University of Hawaii in 1970.

that calls can be correctly routed. In mobile satellite networks, this may be network-based or terminal-based. Terminal-based solutions typically rely on signals from the Global Positioning System (GPS); an accurate approach that has the downside of only working when several GPS satellite signals can be received. Network-based solutions rely on the estimation of location on the basis of timing, arrival angle, and signal strength from one or more base stations. This approach provides less accurate position information. Both of these approaches may fail in dense urban canyons. The need for precise terminal location information depends on the amount that this information is used by handoff and routing algorithms to predict the future evolution of the network topology. Of course, satellite position information can be known accurately due to the predictable movement of the satellites and telemetry, tracking, and control (TTC) communications links with ground stations.

4.1.1. Handoffs. Handoff management is a key determinant in the performance of LEO satellite networks. We have already discussed some of the issues regarding handoff. In general, handoffs should be optimized (which in many situations means minimized), since they are typically accompanied by signaling and call processing overhead, and may result in a degradation in the call's quality of service, due to either blocking or suboptimal routing.

There are two types of handoff in a LEO satellite network. ISL handoffs are generally regular and predictable events. Terminal-to-satellite link-handoffs are not predictable with as much accuracy. Terminals may initiate handoffs to new spot beams or satellites by monitoring signal strength and interference from different carriers. When the terminal enters an overlap area, it requests a handoff. Conversely, in some systems such as Iridium, a gateway station may control the handoff between two satellites by instructing the leading satellite to prepare to handoff and the trailing satellite to prepare to accept the call [18].

Links that must be transferred from one beam to another are subject to blocking if insufficient resources are available in the new beam. Two common techniques for minimizing the possibility that an active call is dropped are to implement guard channels or to queue handover requests. Guard channels are a pool of channels explicitly reserved for handoffs (i.e., no new call arrivals can be assigned a guard channel). Queueing techniques prescribe that when the terminal is in an overlap area that it hold onto its existing channel until a channel in the new beam becomes available. On release of a channel, queued handover requests are served before new calls are admitted. Both approaches lead to fewer handover failures at the expense of a higher initial call blocking probability.

More sophisticated handoff techniques attempt to exploit predictable aspects of the constellation evolution. For example, if it can be estimated when calls will need to be handed off, active channels in one beam can be reserved to handle channels that are predicted to need handoff from another beam in the future [19]. If terminal locations are precisely known, connection admission control can be optimized by predicting the future spotbeam handoff path of each new call arrival [20].

Finally, if ISLs are not used, the use of CDMA for multiple access can yield link performance gains during handoffs. In the Globalstar system, since the same gateway station is used before and after intersatellite handoff, the so-called soft handoff technique of CDMA can be implemented. In this technique, the mobile terminal signal is passed through both the old and new satellites, and the two signals are independently demodulated, selected, and constructively combined to yield processing gain at the edges of satellite coverage.

4.1.2. Multihop Satellite Routing. Consider the establishment of a satellite-based connection traversing ISLs. This requires that a candidate route be picked, each node be signaled about the new connection, the connection be maintained even in the face of changes to the topology, and the connection resources be released once the call is completed. Note that, as exemplified by Fig. 4, there may be many similar routes (topologically) between distant stations. The process whereby a multihop satellite route is established and maintained is a routing problem.

First, consider the establishment of the initial route. In a traditional network, a shortest-path algorithm, perhaps weighted by the amount of current congestion at the nodes, would be used. In the satellite case, the additional consideration of *link permanence* can be applied. In this scenario, routes can be avoided for which it is likely that one of the constituent links is known to be short-lived (such as an ISL about to be deactivated or handed off). In order to minimize handoffs, researchers have studied techniques that consider the time-varying topology during route selection, favoring routes requiring fewer handoffs [21–23]. Another consideration that may be included in routing decisions involves accounting for nonuniform traffic densities in different areas [24,25].

Next, consider a terminal to satellite handoff, which are frequent in LEO constellations. If an ISL exists between the previous satellite and the new satellite, then this link can be grafted onto the existing route without disrupting the other nodes along the path [26]. Note, however, that the new route may no longer be optimal, and over time, may become grossly distorted. Satellite constellations, therefore, may consider this *route augmentation* as a preferred option so long as the resultant route does not fall below some threshold of optimality.

Finally, consider a topological change in which the route from ingress satellite to egress satellite cannot be maintained, or in which the augmented route becomes too suboptimal. In these cases, it may be necessary to determine a new route altogether, and inform the affected nodes along the paths to synchronously switch over at some time instant.

Note that these topological changes can cause the overall circuit delay to drastically change, which may be a problem for some services. One way to compensate for this is to use buffers at the endpoints of the satellite connections to smooth out any delay variations, at the expense of consistently larger delays.

4.2. Packet-Switched Architectures

An alternative to circuit switching, especially suited for interworking with actual packet-switched networks like the Internet, is a packet-switched satellite network. This approach has the advantage of not requiring per connection state to be kept and maintained on board the spacecraft. Nevertheless, many of the same handoff challenges described above still persist, because for reasons of link efficiency, channel reservations between terminals and satellites are still desirable.

There are several different techniques available for implementing packet routing in satellite networks, differing chiefly in their implementation and processing complexity onboard the satellites. A general discussion of several IP networking issues, including address translation, multicast, interfacing with exterior routing protocols, tunneling, and quality of service can be found in Ref. 27. In this section, we focus on the basic packet routing problem in a satellite constellation.

Consider the problem of routing a packet from one terminal to another through one or more satellite nodes. The simplest approach to route the packet from the standpoint of satellite complexity would be to flood the packet (i.e., transmit the packet out of all the active link interfaces except the one on which the packet arrived) and limit the number of hops for which the packet can be forwarded (such as by decrementing a counter). Because of the densely interconnected mesh, such an approach would be grossly suboptimal, leading to extremely congested networks. Another simple approach from the satellite standpoint would be to determine the entire route of the packet a priori at the ingress terminal, and affix this route to the packet before sending it to the first satellite node. Each satellite would then forward the packet by simply following the next-hop instructions attached to the packet, and an onboard routing table would not need to be maintained. This would require, however, that the terminals affixing the route to the packet determine the optimal route; that is, they must have access to full instantaneous routing state of the network. The burden placed on terminals on the network may be considerable in this case. An alternative could be to have route servers distributed throughout the network that could be queried by terminals whenever a new route was needed. However, this approach would incur extra latency in the initiation of the communications. A more serious impediment to this approach would occur if route topological changes were not predicted (such as a terminal-initiated handoff). In this case, packets could be lost to a deadend until new route information is made available to endpoints, and because of the latency in the system, it would take some time for the routing information to stabilize on any unanticipated topological change.

To offload the responsibility of routing from terminals to the satellite network, again, different approaches may be used. For example, centralized routing servers could periodically upload routing tables to satellites, incrementally updating them as topology changes become known. Again, there may be latency issues with this approach upon unexpected topological changes. This approach, while requiring the satellite to maintain memory for and lookup routes from a routing table, the task of actually computing the routing tables is left to the ground segment.

Latency issues in the propagation of state information can be minimized if the satellite nodes implement fully distributed routing, such as used in the Internet. The drawback to this approach is that it requires satellites to not only build and maintain routing tables but also incur the processing and signaling overhead of a distributed routing protocol. Indeed, terrestrial Internet routing protocols such as traditional distance vector or link-state protocols, applied to this type of a dynamic network topology, would either be slow to converge or would overwhelm nodes with update messages. General flooding of routing update messages would also be problematic, even if there were sufficient ISL capacity to handle the messages, due to the sheer volume of routing updates that would need to be processed. Nevertheless, distributed routing techniques have been the focus of most recent research, as researchers have studied ways to simplify the problem by exploiting the regularity of the network topology and the predictability of ISL topological changes.

One general technique for simplifying distributed routing is to try to hide the mobility of satellite nodes from the terrestrial nodes. The semiregular structure of most satellite constellations facilitates this. For example, if one overlays a cellular structure over the earth's surface, with the cell size roughly corresponding to the coverage area of a satellite footprint, then it may be possible to overlay a logical network structure of "virtual nodes," in which different satellites over time embody each virtual node [28]. Another possible approach is to assume that the satellite network evolves through a finite series of topologies, and have the satellite network store the appropriate routing table for each state and iterate through these tables [29]. Although such approaches appear to have some promise when considering idealized constellation geometries, they have yet to be demonstrated as a robust approach when applied to practical constellations [30].

5. FUTURE DIRECTIONS

The design of a satellite constellation is a complex optimization problem with the cost a function of various link parameters as well as terminal and satellite complexity. In this article we have provided an overview of nongeostationary satellite fundamentals and surveyed many of the design features that differentiate these networks from systems based on geostationary satellites. Unlike GSO systems, LEO and MEO satellite networks are still in their infancy, and several of the initial attempts to deploy large-scale commercial constellations have been a financial failure. Nevertheless, the promises of global ubiquitous coverage, accompanied by significantly lower propagation delays, will continue to spur development of nongeostationary satellite network architectures. Technically, many issues will be the subject of ongoing research and development, including interference mitigation, link issues (such as error control coding and handoff algorithms), electronically steerable antennas,

routing algorithms, onboard switching architectures, electronics based on more radiation-resistant substrates (such as gallium arsenide), and regulatory issues.

BIOGRAPHY

Thomas R. Henderson received his B.S. and M.S. degrees from Stanford University, Stanford, California, and a Ph.D. from the University of California, Berkeley, California, all in electrical engineering. He is currently a researcher at Boeing Phantom Works, the research and development division of The Boeing Company. He is also presently a part-time lecturer in the Electrical Engineering department at the University of Washington, Seattle, Washington. Previously, he was director of digital television research and standards at Geocast Network Systems in Menlo Park, California. Prior to attending Berkeley, he worked at COMSAT Laboratories in Maryland, where his responsibilities included performance analysis, protocol development, and standards activities in the areas of ATM and ISDN over satellite networks. He has been a rapporteur of ITU-T Study Group 13, a vice chair of ANSI T1S1.5, and editor of several national and international standards. His current research interests are focused on network-layer mobility and routing for wireless IP networks.

BIBLIOGRAPHY

1. S. R. Pratt, R. E. Raines, C. E. Fossa Jr., and M. A. Temple, An operational and performance overview of the IRIDIUM low Earth orbit satellite system, *IEEE Commun. Surv.* Second Quarter: **2**(2): 2–8 (1999).
2. E. Hirshfield, The Globalstar system: Breakthroughs in efficiency in microwave and signal processing technology, *Space Commun.* **14**: 69–82 (1996).
3. B. Pattan, *Satellite-Based Cellular Communications*, McGraw-Hill, New York, 1998.
4. G. Maral, J.-J. De Ridder, B. G. Evans, and M. Richharia, Low Earth orbit satellite systems for communications, *Int. J. Satellite Commun.* **9**: 209–225 (1991).
5. G. Maral and M. Bousquet, *Satellite Communications Systems*, Wiley, Chichester, UK, 2000.
6. L. Ghedia, K. Smith, and G. Titzer, Satellite PCN — the ICO system, *Int. J. Satellite Commun.* **17**: 273–289 (1999).
7. J. Walker, Some circular orbit patterns providing continuous whole earth coverage, *J. Br. Interplan. Soc.* **24**: 369–381 (1971).
8. R. Thurman and P. Worfolk (No date), SaVi (online), *http://www.geom.umn.edu/worfolk/SaVi/*, April 4, 2001.
9. L. Wood, *Internetworking with Satellite Constellations*, Ph.D. thesis, Univ. Surrey, 2001.
10. J. E. Draim, Design philosophy for the ELLIPSO™ satellite system, *Proc. 17th AIAA Int. Comm. Satellite Conf.*, Yokohama, Japan, 1998.
11. I. Ali, N. Al-Dhahir, and J. E. Hershey, Doppler characterization for LEO satellites, *IEEE Trans. Commun.* **46**: 309–313 (1998).
12. F. Vatalaro and G. E. Corazza, Probability of error and outage in a Rice-lognormal channel for terrestrial and satellite personal communications, *IEEE Trans. Commun.* **44**: 921–924 (1996).
13. G. E. Corazza and F. Vatalaro, A statistical model for land mobile satellite channels and its application to nongeostationary orbit systems, *IEEE Trans. Vehic. Technol.* **43**: 738–742 (1994).
14. G. E. Corazza and C. Caini, Satellite diversity exploitation in mobile satellite CDMA systems, *Proc. Wireless Comm. Networking Conf.* (*WCNC*), 1999, pp. 1203–1207.
15. J. Restrepo and G. Maral, Cellular geometry for world-wide coverage by non-GEO satellites using "Earth-fixed cell" technique, *Space Commun.* **14**: 179–189 (1996).
16. M. Werner, A. Jahn, E. Lutz, and A. Bottcher, Analysis of system parameters for LEO/ICO-satellite communication networks, *IEEE J. Select. Areas Commun.* **13**: 371–381 (Feb. 1995).
17. F. Ananasso and M. Carosi, Architecture and networking issues in satellite systems for personal communications, *Int. J. Satellite Commun.* **12**: 33–44 (1994).
18. Y. C. Hubbel, A comparison of the IRIDIUM and AMPS Systems, *IEEE Network Mag.* **11**: 52–59 (1997).
19. P. Wan, V. Nguyen, and H. Bai, Advance handovers arrangement and channel allocation in LEO satellite systems, *Proc. IEEE Globecom*, 1999, pp. 286–290.
20. S. R. Cho, I. F. Akyildiz, M. D. Bender, and H. Uzunalioglu, A new spotbeam handover management technique for LEO satellite networks, *Proc. IEEE Globecom*, 2000, pp. 1156–1160.
21. M. Werner et al., ATM-based routing in LEO/MEO satellite networks with intersatellite links, *IEEE J. Select. Areas Commun.* **15**: 69–82 (1997).
22. A. Jukan, H. N. Nguyen, and G. Franzl, QoS-based routing methods for multihop LEO satellite networks, *Proc. IEEE Int. Conf. Networks* (*ICON*), 2000, pp. 399–405.
23. H. Uzunalioglu, Probabilistic routing protocol for low earth orbit satellite networks, *Proc. IEEE Int. Conf. Commun.* (*ICC*), 1998, pp. 89–93.
24. A. Jamalipour, *Low Earth Orbital Satellites for Personal Communication Networks*, Artech, Norwood, MA, 1998.
25. Y. Kim and W. Park, Adaptive routing in LEO satellite networks, *IEEE Vehicular Tech. Conf.*, 2000, pp. 1983–1987.
26. H. Uzunalioglu and W. Yen, Managing connection handover in satellite networks, *Proc. ACM Mobicom*, 1997, pp. 204–214.
27. L. Wood et al., IP routing issues in satellite constellation networks, *Int. J. Satellite Commun.* **19**: 69–92 (2001).
28. R. Mauger and C. Rosenberg, QoS guarantees for multimedia services on a TDMA-based satellite network, *IEEE Commun. Mag.* **35**: 56–65 (1997).
29. H. S. Chang et al., Topological design and routing for LEO satellite networks, *Proc. IEEE Globecom*, 1995, pp. 529–535.
30. T. R. Henderson and R. H. Katz, On distributed, geographic-based packet routing for LEO satellite networks, *Proc. IEEE Globecom*, 2000, pp. 1119–1123.
31. L. Wood (n.d.), Lloyd's satellite constellations (Online), *http://www.ee.surrey.ac.uk/Personal/L.Wood/*, April 4, 2001.

LINEAR ANTENNAS

SHELDON S. SANDLER
Lexington, Massachusetts

An antenna constructed of a few straight-line segments, made of conducting, partially conducting, or nonconducting material, is known as a *linear antenna*. Because of their simplicity, linear antennas are probably the most common type of radiator for communication between distant points. They exist in many varieties delineated by (1) geometry (e.g., straight dipole, V-shaped dipole, L-shaped antenna), (2) electrical characteristics (e.g., resonant, antiresonant, wideband), and (3) radiation properties (e.g., isotropic, directive). With a view toward application, linear antennas and antennas in general are evaluated with respect to the spatial and frequency characteristics of the radiation and their circuit or electrical properties. For example, if a designer wants to send narrowband signals to all parts of the world without any preference in direction, the ideal radiator would have an isotropic distribution of energy in space. Furthermore, as a circuit element the antenna must be matched to a low-impedance source through a transmission line. Here a resonant antenna is the right choice.

1. LINEAR DIPOLES

To better understand the linear antenna, it is best to start out with the simplest example, namely, a linear dipole of half-length h driven in the center by a sinusoidal voltage. The dipole is constructed of thin wire or rods, with each rod being connected to one end of the transmission line, as shown in Fig. 1.

The dipole of Fig. 1 has radiation characteristics that are dependent on the current on the wires. A good analogy in visualizing the current distribution is to consider a string fixed at both ends and plucked at various points along the length. After plucking, nodes and antinodes appear in a configuration called a *standing wave*. For our dipole, the first "mode" corresponding to a resonant length resembles a cosine with a maximum at the center (antinode) and zero at the ends (nodes) as shown in Fig. 2. This would be a half-wavelength dipole, $h/\lambda = 0.25$. If the dipole is electrically smaller, say, $h/\lambda = 0.05$, the current would still be zero at the ends and the maximum would still be at the center. As the electrical length of the antenna increases, by increasing the frequency of the source, more complicated current distributions arise. In fact, with our simple model, there will be times when the driving point is at a node, and the driving point impedance is infinite. This scenario is shown in Fig. 2 for $h/\lambda = 0.5$.

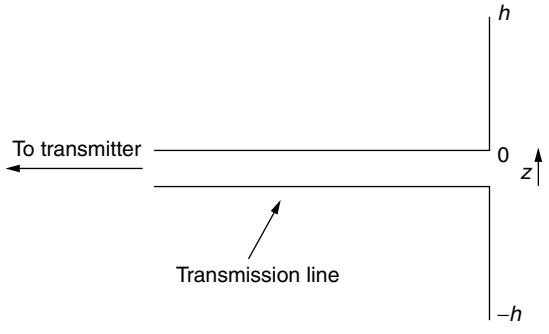


Figure 1. A dipole antenna.

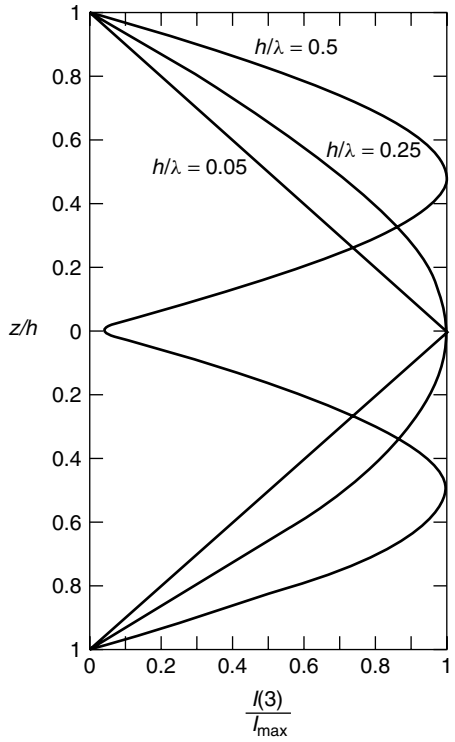


Figure 2. Idealized antenna currents.

In practice the current is never a pure sinusoid, so that the driving-point impedance will never be infinite. The spatial radiation characteristics, called the *radiation pattern*, after calculation from an assumed current, have their maximum perpendicular to the dipole when the half-length is less than about a wavelength and a half. This maximum is in the plane of the antenna (i.e., E plane) in the broadside direction and in the plane perpendicular to the antenna (i.e., H plane). The radiation is uniform or isotropic. Figure 3 shows the E plane radiation pattern for a linear antenna that has values of $h/\lambda = 0.05$, 0.25, and 0.5. For increasing lengths, the maximum radiation can be in oblique directions to the dipole axis. It is instructive to quantitatively examine the radiation pattern of a linear antenna based on the geometry shown in Fig. 4. The far-zone electric field E_θ^R, also called the *radiation field*, is in the direction of the θ arrow and is tangent to a sphere whose radius is R. Together, R and θ form the E plane. Note that one must be roughly in the range $k_0R \gg 1$ to meet the conditions for the far zone. A simplified relation for E_θ^R is

$$E_\theta^R = CF_0(\theta, k_0h) \tag{1}$$

where C is a factor that contains the R dependence of the field, and F_0 is the field pattern normalized with respect to the value of the current at $z = 0$.

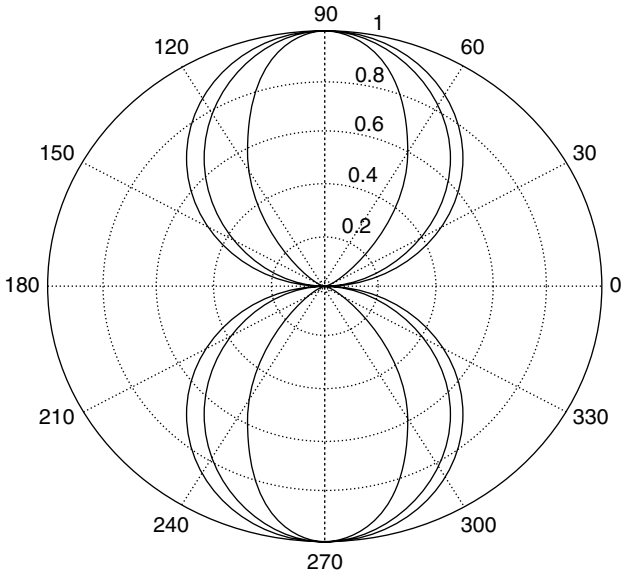


Figure 3. Radiation pattern for linear antenna with different electrical lengths.

The field pattern is given by [1]

$$F_0(\theta, k_0 h) = \frac{k_0 \sin\theta}{2I(0)} \int_{-h}^{h} I(z') e^{jk_0 z' \cos\theta} dz' \qquad (2)$$

When the current at the base of the antenna is zero (see Fig. 2 for $h/\lambda = 0.5$), the field pattern can be normalized to the maximum value of the current. This produces a radiation pattern $F_m(\theta, k_0 h)$. Patterns shown in Fig. 3 were computed from the F_0 or F_m relation and are valid for the E plane. The far-zone magnetic field E_ϕ^R is orthogonal to the E_θ^r field and is located in the H plane formed by the ϕ and r arrows in Fig. 4.

The antenna current, which in the frequency domain is a complex quantity, completely determines the driving-point impedance of the antenna. For an electrically short antenna, the driving-point resistance is small and the reactance is capacitive and large. At the first resonance, the complex driving-point impedance is approximately

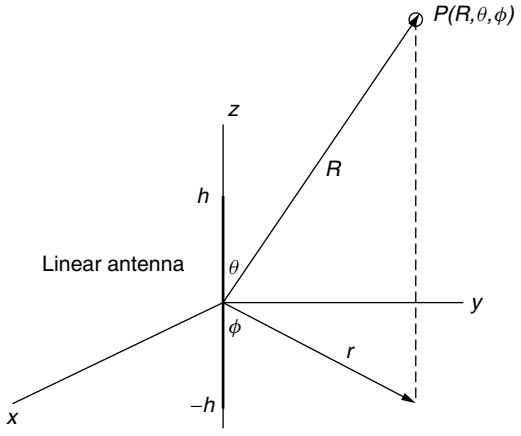


Figure 4. Gain and directivity.

$73 + j42$ ohms (Ω) (half-wavelength antenna). This explains why the half-wavelength linear antenna is so popular, since the driving-point resistance is easy to match with available transmitters and transmission lines. Many sources giving the impedance characteristics of linear antennas are available; perhaps the best is in *Tables of Antenna Characteristics* by King [2]. The design of a linear antenna system for a single frequency is not very complicated. Choose a half-wavelength antenna, design a matching network to cancel out the driving-point reactance, and find a compatible transmitter and coaxial line to match the antenna. The design of a linear antenna system where the bandwidth is important requires that the response of the antenna at different frequencies not cause excessive degradation in the amplitude of the transmitted signal.

One important antenna parameter has to do with the concentration of radiation in a specific direction. The gain of an antenna is proportional to the power radiated in a given direction divided by the average power. Another parameter, called the *directivity*, is equal to the maximum value of the gain and is expressed as a numeric or in decibels (dB). The directivity of a half-wavelength dipole is 1.64 or 2.1 dB. (i.e., $10 \log 1.64$). An example to illustrate these concepts is to design a more directive linear antenna that carries a sinusoidal current. In analogy to the radiation (light pattern) from a thin optical slit, it is known that a more directive light pattern is obtained by increasing the length of the slit. However, when the same idea is tried with a linear antenna, an increase in directivity is not present when the antenna is lengthened. This is because successive half-wavelengths of current are of opposite sign and serve to reduce the radiated field. To overcome this difficulty, phase-reversing stubs can be placed every half-wavelength along the antenna. The current along the antenna is now unidirectional and a closer approximation to the uniform light in a slit.

Sometimes the designer is limited in the physical length available for the antenna. For example, linear antennas that are short in electrical length can have reduced resistance and increased capacitance when compared with a half-wavelength dipole. The driving-point impedance of a quarter-wavelength dipole is about $14 - j195$ Ω, while the impedance of a half wavelength dipole is about $73 + j42$ Ω. To make increase the apparent electrical length of the quarter-wavelength dipole, a series inductance can be placed near the base of the antenna, say, with a coil of wire. Top loading and series loading at any point is also possible to change the current distribution on the antenna.

2. TRAVELING-WAVE ANTENNAS

So far, the discussion has been concerned mainly with the standing-wave linear antenna, since the current must be zero at the ends of the antenna (i.e., $z = \pm h$). A different type of antenna current distribution is concerned with traveling waves instead of standing waves. It is well known that a standing wave can be decomposed into a forward/backward-traveling wave. For example, in a half

wavelength dipole the ideal current is given by $I_z(z) = I_0 \cos k_0 z$, $|z| \le h$. Using the exponential representation for the current, we obtain

$$I_z = I_0 \cos k_0 z = I_0 \left(\frac{e^{jk_0 z} - e^{-jk_0 z}}{2} \right) \tag{3}$$

Using $e^{j\omega t}$ time dependence,

$$I_z = I_0 e^{j(\omega t + k_0 z)} + I_0 e^{j(\omega t - k_0 z)} \tag{4}$$

where $k_0 = (2\pi/\lambda_0)$ = free-space propagation constant
$\omega = 2\pi f$
f = frequency in Hz
t = time
z = distance along the antenna (5)

The first term on the right represents a traveling wave moving inward to the base, and the second term represents a wave moving outward toward the end of the antenna at $z = h$.

From transmission-line theory it is also known that using a termination equal to the characteristic impedance can produce a reflectionless line. A traveling-wave antenna, called a "beverage antenna," is constructed by placing a conductor parallel to the earth and terminating it with terminal impedance, producing minimum reflections at the end.

For a monopole structure the radiation pattern in the E plane roughly resembles a set of rabbit ears, where each ear is at an oblique angle to the antenna axis. As the monopole elongates electrically, the ears move closer to the antenna axis. The radiation pattern of a traveling-wave dipole antenna consists of two rabbit ears roughly in the shape of the letter X. If a unidirectional pattern is desired, a V-shaped antenna may be used. It is constructed by bending the arms of a dipole about the center. The apex angle is chosen such that the inside radiation lobes are completely superimposed on one another. A diagram of the radiation patterns for two representative traveling wave antennas is shown in Fig. 5.

The reflectionless antenna described by Wu and King [3] has a prescribed resistive coating that produces a traveling wave along a finite-length antenna. It has an important application as an antenna for pulses that have a very wideband frequency spectrum. One major drawback for this antenna is that it is about 50% efficient since half of the power is dissipated in the resistance. Increased attention has been given to antennas energized by short temporal pulses for use in GPR (ground-penetrating radar) systems. The design of antennas for use in such systems involves somewhat different criteria and physical viewpoint than antennas used for CW (continuous-wave) systems. The following simple example will illustrate the physics involved in driving a linear antenna with a carrierless temporal pulse. A dipole of finite length is energized with a short temporal pulse, say, with a half-width of an nanosecond (1 ns). In order to get the whole pulse to exist on the antenna, the antenna length must be greater than a foot. Roughly, the pulse travels at a velocity of 1 ns/ft along the antenna. This is the velocity of an electromagnetic wave in free space.

2.1. Example of a Traveling-Wave Antenna

To gain some insight into the radiation properties of a pulsed antenna, the traveling-wave antenna with a Gaussian pulse excitation will be considered. A more detailed analysis can be found in the book by Smith [4]. The radiated field for a linear antenna in the time domain has a form different from that in the frequency-domain integral given earlier. It is possible to find the time-domain radiation expression from the frequency-domain representation by using a Fourier transform. Qualitatively the time-domain radiated field is proportional to the integral along the antenna of the time derivative of the current, evaluated in retarded time. Expressed in quantitative terms, this radiated field for a monopole is given by

$$E^r = \frac{\mu_0}{4\pi} \sin\theta \int_0^h \left[\frac{\partial}{\partial t'} I_z \left(t' - \frac{z'}{c} \right) \right] dz' \tag{6}$$

where

μ_0 = magnetic constant
$= 4\pi \times 10^{-7}$ H/m (henries per meter) (7)
= retarded time

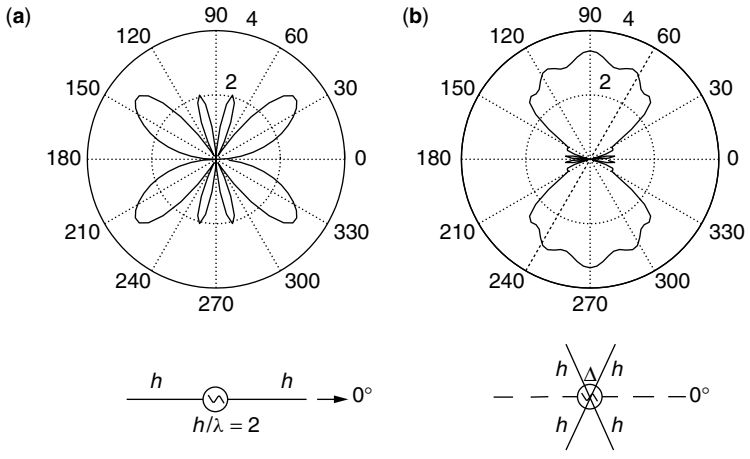


Figure 5. Two representative traveling wave antennas: (**a**) linear antenna with an assumed traveling wave current ($h/\lambda = 2.0$); (**b**) X-shaped antenna with traveling-wave currents ($\Delta = 32°$, $\Delta/\lambda = 2$).

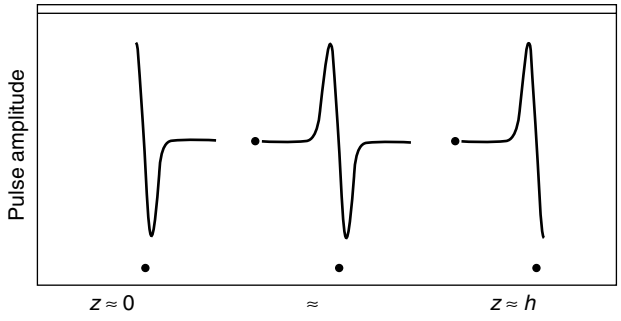


Figure 6. Pulse example.

In expression (6) the current is traveling along the positive z direction with a velocity c. Figure 6 shows the integrand of the radiation integral of (6). The first derivative of a Gaussian pulse has two equal sections, one positive and one negative, and at the driving point there are times (i.e., near $z = 0$) that areas under the positive and negative sections do not cancel. Here radiation exists. When the entire pulse is present in the antenna, say, near $z = h/2$, the positive and negative areas do cancel and there is no radiation. When the pulse is near the end of the antenna, $z = h$, the incident and reflected pulse areas may not cancel at certain times, giving rise to a second radiated pulse. As time progresses, pulses continue to be radiated at $z = 0$ and $z = h$.

2.2. Receiving Antenna

Linear antennas are also used as receptors for electromagnetic signals. Important quantities are the voltage at the terminals of the antenna, V_0, and the current through the load impedance, Z_L. A sketch of a linear antenna with length $2h$ used for reception is shown in Fig.7, along with its Thévinen equivalent circuit. The equivalent circuit has a series arrangement of the driving-point impedances Z_0 and Z_L driven by the open-circuit voltage V_0. From this arrangement the current in the circuit is given by

$$I_z(z = 0) = I_z(0) = \frac{V_0}{Z_0 + Z_L} \tag{8}$$

An important parameter for the receiving antenna is the complex effective length $h_e(k_0 h)$, which relates V_0 to the antenna. Thus

$$V_0 = 2h_e(k_0 h)E_z^{\text{inc}} \text{ volts } (V) \tag{9}$$

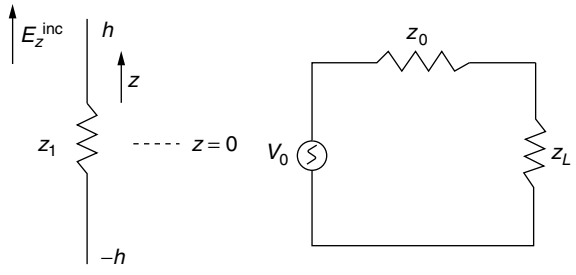


Figure 7. Receiving antenna and its Thévenin equivalent circuit.

A good reference on the receiving antenna is the book by King [5]. When the antenna is short, $k_0 h \leq 0.5$, the effective length is approximately equal to half the physical length. Stated in another way, the total length of the antenna is about equal to twice the effective length. As the antenna becomes longer than $k_0 h = 0.5$, the simple approximation breaks down. For example, a resonant antenna with $k_0 h = \pi/2$ has an effective length of about $h_e \approx 1.21$. So far the discussion has been about the circuit properties of transmitting and receiving antennas. When attention is given to the currents on transmitting and receiving antennas, the situation is more complicated. These currents differ because on a receiving antenna, both the incident electric field and the load impedance are involved. With a zero value of Z_L, the current distribution is of the receiving type. For antennas of moderate length the current has a shifted cosine distribution given by $\cos k_0 z - \cos k_0 h$. When the load impedance is increased, a transmitting current is added to the receiving current. It has the form $\sin k_0(h - |z|)$.

BIOGRAPHY

Sheldon S. Sandler received his Ph.D. and M.A. from Harvard University, Massachusetts, his M.Eng.E.E. from Yale University, and his B.S.E.E. from Case Institute of Technology. He is a professor emeritus in the Department of Electrical and Computer Engineering at Northeastern University, Boston, Massachusetts. He is also a research fellow in the Department of Archaeology at Boston University, Massachusetts, working on remote sensing for site evaluation. At Geo-Centers, Inc., Massachusetts, he is a senior engineer and a technical advisor to the CEO. There he has developed new GPR systems and time domain antennas as well as designing algorithms to detect targets in noisy data. Dr. Sandler has been a guest professor at both the E.T.H and the University of Zurich in Switzerland and at the Robotics Center at the University of Rhode Island. At the MRC Laboratory in Cambridge, England, he was a visiting scholar. Dr. Sandler is the author of *Picture Processing and Reconstruction* and the coauthor of *Arrays of Cylindrical Dipoles* with R.W.P. King and R. Mack. He is a member of the International Radio Union (URSI), IEEE, Sigma Xi, Tau Beta Pi, and Eta Kappa Nu.

BIBLIOGRAPHY

1. R. W. P. King, R. B. Mack, and S. S. Sandler, *Arrays of Cylindrical Dipoles*, Cambridge Univ. Press, Cambridge, UK, 1968.
2. R. W. P. King, *Tables of Antenna Characteristics*, IFI/Plenum, New York, 1971.
3. T. T. Wu and R. W. P. King, The cylindrical antenna with nonreflecting resistive loading, *IEEE Trans. Antennas Propag.* **AP-13**: (1975).
4. G. S. Smith, *An Introduction to Classical Electromagnetic Radiation*, Cambridge Univ. Press, Cambridge, UK, 1997, Chap. 8.
5. R. W. P. King, *The Theory of Linear Antennas*, Harvard Univ. Press, Cambridge, MA, 1956, Chap. 4.

LINEAR PREDICTIVE CODING

AMRO EL-JAROUDI
University of Pittsburgh
Pittsburgh, Pennsylvania

1. INTRODUCTION

Most real-world signals carry redundant information from one sample (or snapshot) to the next. For example, a television video signal is made of a sequence of frames (about 30 frames per second, depending on the video standard) where often very little changes in the picture from one frame to the next. Even within a frame, neighboring pixels are likely to be related in terms of intensity and color. It is not unusual for an office document to contain long strings of consecutive white pixels or long strings of consecutive black pixels.

From an efficient communication standpoint, it is extremely wasteful to spend valuable bits (or bandwidth) on encoding the redundant information from one sample to the next. Instead, it is more efficient to use the bits to encode only the novel information. Consequently, a preprocessing procedure to remove the intersample redundancy becomes necessary before encoding a signal for storage or transmission. This procedure has two steps. In the first step, an estimate of the current sample is *predicted* (guessed scientifically) based on its neighbor(s). This predicted value is the redundant portion of the current sample since it is based solely on neighboring samples. The second step is simply to subtract the predicted value (the redundancy) from the current sample, thereby, leaving only the novel information to be encoded. Although, in general, one may use any parametric function to predict a sample from its neighbors, the discussion below focuses on *linear prediction* (LP), where the sample is predicted as a linear combination of other samples. While the focus of this article is on the use of LP in redundancy removal or data compression for coding applications [also known as *linear predictive coding* (LPC)], it is important to note that LP is used in a variety of other applications, including forecasting, control, system modeling and identification, and spectral estimation, to name only a few [1,2].

The remainder of the article is organized as follows. In Section 2, LP is formulated and the optimal prediction parameters are derived. In Section 3, the computational aspects and algorithms for the implementation of LP are explored. In Section 4, examples and applications of LPC are presented. Finally, in Section 5, variations on LPC used for coding speech signals are discussed.

2. FORMULATION OF LINEAR PREDICTION

Given a discrete-time signal,[1] x_n, defined over a finite interval $n = 0, 1, 2, \ldots, N-1$, define $\hat{x}_n$ to be the predicted value of x_n based on the p previous values of x_n. In other words

$$\hat{x}_n = \sum_{k=1}^{p} a_k x_{n-k} \tag{1}$$

where $a_1, a_2, \ldots, a_p$ are the prediction parameters (or coefficients). The prediction error e_n is then defined as the difference between x_n and $\hat{x}_n$:

$$e_n = x_n - \hat{x}_n = x_n - \sum_{k=1}^{p} a_k x_{n-k} \tag{2}$$

The error signal is often referred to as the *residual signal* since it describes the residual information after the redundancy removal.

The prediction parameters are chosen to minimize the prediction error subject to an optimality criterion. Different prediction parameters can be obtained depending on the criterion selected. Below, we examine the method of *least squares* (LS), which is one of the more popular criteria. For an example of other methods, please see Refs. 3 and 4.

2.1. LS Minimization

The least-squares criterion takes on different forms depending on the assumptions made regarding the signal, x_n. If we treat the signal as a random signal, we minimize the expected value of the squared error

$$\mathbf{E} = E\{e_n^2\} = E\left\{\left[x_n - \sum_{k=1}^{p} a_k x_{n-k}\right]^2\right\} \tag{3}$$

where $E\{\cdot\}$ stands for the expectation operator. Assuming the signal to be a sample of a stationary process, substituting for e_n in Eq. (3), taking the derivative with respect to a_i for $1 \le i \le p$, and setting the derivative to zero yields the following set of equations

$$\sum_{k=1}^{p} a_k R_{i-k} = R_i \tag{4}$$

where R_i is the autocorrelation of the random process and is defined as

$$R_i = E\{x_n x_{n-i}\} \tag{5}$$

Here, R_i measures how a sample is related to another sample which is i lags away, with a value of zero indicating no correlation between the samples. In other words, the autocorrelation function is a measure of the average redundancy between samples. It is then no surprise that the autocorrelation information is used to determine the optimal prediction parameters.

Instead of treating x_n as a random signal, we may treat it as a deterministic signal. Then, we minimize the sum of the squared errors

$$\mathbf{E} = \frac{1}{N}\sum_n e_n^2 = \frac{1}{N}\sum_n \left[x_n - \sum_{k=1}^{p} a_k x_{n-k}\right]^2 \tag{6}$$

[1] A discrete-time signal is usually obtained by sampling an analog signal using an analog-to-digital converter.

If we assume that the range of minimization is infinite, then taking the derivative and setting it to zero yields

$$\sum_{k=1}^{p} a_k \hat{R}_{i-k} = \hat{R}_i \tag{7}$$

where $\hat{R}_i$ is the (time-average) autocorrelation function of the signal x_n and is given by

$$\hat{R}_i = \frac{1}{N} \sum_{n=-\infty}^{\infty} x_n x_{n-i} \tag{8}$$

Note that $\hat{R}_i$ here is calculated over all the values of x_n. Unfortunately, in practice, the signal x_n is not given over an infinite interval. To reduce the range of summation in Eq. (8), the given signal is often multiplied by a window function creating a new signal $\tilde{x}_n$, which is zero outside the interval $0 \le n \le N-1$. The autocorrelation function is then calculated for $\tilde{x}_n$,

$$\hat{R}_i = \frac{1}{N} \sum_{n=0}^{N-1} \tilde{x}_n \tilde{x}_{n-i} \tag{9}$$

where, as described above, $\tilde{x}_n$ is given by

$$\tilde{x}_n = \begin{cases} x_n w_n, & 0 \le n \le N-1 \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

Clearly the window function, w_n, will influence the value of the autocorrelation function and the resulting predictor coefficients. Consequently, great care should be used in selecting it. Two of the more popular window functions are the rectangular window given by $w_n = 1$ for $0 \le n \le N-1$, and the Hamming window given by

$$w_n = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \qquad \text{for} \qquad 0 \le n \le N-1 \tag{11}$$

It is important to note the similarity between Eqs. (4) and (7). They differ only in the definition of the autocorrelation function. Fortunately, in practice, (9) is often used to estimate the autocorrelation of the stationary process. In this case, the predictor coefficients produced under the random stationary process assumption would be the same as those produced under the deterministic infinite interval assumptions. Since the methods rely on the autocorrelation information, they are often referred to as the *autocorrelation method of LP*.

If we assume that the error in Eq. (6) is defined over a the finite interval $0 \le n \le N-1$ and minimize it only over this interval, we obtain the following set of equations

$$\sum_{k=1}^{p} a_k \varphi_{i,k} = \varphi_{0,i} \tag{12}$$

where $\varphi_{i,k}$ is called the *covariance of the signal* x_n and is given by

$$\varphi_{i,k} = \frac{1}{N} \sum_{n=0}^{N-1} x_{n-i} x_{n-k} \tag{13}$$

In Eq. (13), it is required that the values of the signal x_n be known over the range $-p \le n \le N-1$ for all the terms in the summation to be calculated. If the values for $-p \le n \le -1$ are not known, then the summation limits in (13) must be changed to $p \le n \le N-1$. This method is often referred to as the *covariance method of LP*. For the details and properties of this method, please see the article by Makhoul [1].

2.2. The Minimum Error

In order to gauge the quality of the obtained predictor, one can examine the final prediction error which is the minimum value of the error criterion used. This value is, of course, dependent on the method of LP used. For the autocorrelation method, the final prediction error is obtained by substituting Eq. (7) in (3) or (6), thereby producing

$$\mathbf{E}_{\min} = R_0 - \sum_{k=1}^{p} a_k R_k \tag{14}$$

This error is a measure of the variance (or power) in the residual. Consequently, this value is used to calculate a factor, G, to normalize the residual signal and produce a unit-variance excitation signal, u_n. In other words

$$u_n = \frac{e_n}{G} \tag{15}$$

where

$$G = \sqrt{\mathbf{E}_{\min}} \tag{16}$$

The advantage of this normalization lies in that, independent of the power of the original signal, the resulting excitation signal will be consistently unit-variance making it easier to encode. Figure 1 is a block diagram of the relation between the given signal x_n, the error signal e_n, and the excitation signal u_n. It is easy to see that the relation is that of a discrete-time moving-average (MA) linear time-invariant filter with input x_n, output u_n, and parameters $\{1, -a_1, -a_2, \ldots, -a_p\}$ and G. The MA (also known as the *analysis*) filter has a transfer function $A(z)/G$, where

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \tag{17}$$

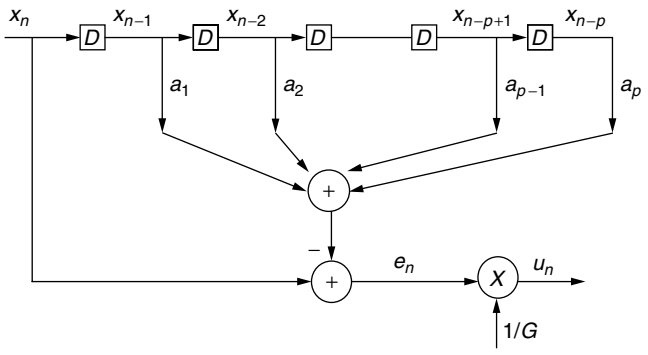


Figure 1. LPC analysis.

As discussed above, the excitation is then encoded for storage or transmission through a communications channel. The prediction parameters and gain are also encoded for storage or transmission. The decoder then uses the residual and the prediction parameters and gain to synthesize the original signal. The synthesis is performed by passing the excitation signal (hence the name) through a filter $H(z)$ that is the inverse of the MA filter used by the encoder. In other words

$$H(z) = \frac{G}{A(z)} \tag{18}$$

and

$$s_n = Gu_n + \sum_{k=1}^{p} a_k s_{n-k} \tag{19}$$

Figure 2 is a block diagram of the synthesis filter that is a discrete-time autoregressive (AR) LTI system. It is important to note that, in Eq. (18), we assumed that the analysis filter $A(z)$ is invertible which is true only if all its roots lie inside the unit circle. This issue will be addressed in a later section. If the encoding of u_n, the prediction coefficients and gain introduced no errors, the synthesized signal s_n would be identical to the original signal x_n. This latter assumption however is not realistic since encoding invariably introduces quantization errors. The effect of quantization on the excitation and prediction parameters will be discussed later.

3. COMPUTATION OF PREDICTION PARAMETERS

In the case of the autocorrelation method, the minimization equations in Eq. (4) form a set of p linear equations in p unknowns, which can be written in matrix form as

$$\mathbf{Ra} = \mathbf{r} \tag{20}$$

where the autocorrelation matrix $\mathbf{R}$ is given by

$$\mathbf{R} = \begin{bmatrix} R_0 & R_1 & R_2 & \cdots & R_{p-1} \\ R_1 & R_0 & R_1 & \cdots & R_{p-2} \\ R_2 & R_1 & R_0 & \cdots & R_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R_{p-1} & R_{p-2} & R_{p-3} & \cdots & R_0 \end{bmatrix} \tag{21}$$

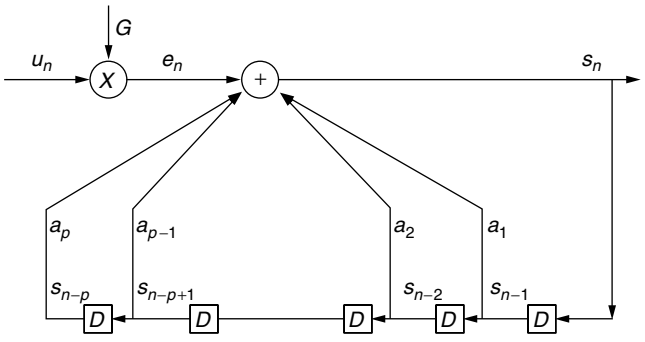


Figure 2. LPC synthesis.

and the vectors $\mathbf{a}$ and $\mathbf{r}$ are given by

$$\mathbf{a} = [a_1, a_2, \ldots, a_p]^T \tag{22}$$

$$\mathbf{r} = [r_1, r_2, \ldots, r_p]^T \tag{23}$$

It is easy to see that $\mathbf{R}$ has a special structure; namely, the elements along each diagonal are equal and the matrix is symmetric. A matrix of this form is called *Toeplitz symmetric*. The vector of prediction coefficients can be obtained by solving Eq. (20) using standard methods such as Gaussian elimination. These methods usually require on the order of p^3 operations. These methods however do not take advantage of the special structure of $\mathbf{R}$. The special method of Levinson–Durbin takes into account the Toeplitz nature of $\mathbf{R}$ and the fact that the $\mathbf{r}$ vector is composed of the same elements in $\mathbf{R}$. This method requires on the order of p^2 operations and is described as follows:

1. The initialization step

$$E_0 = R_0 \tag{24}$$

2. The recursion steps repeated for $i = 1, 2, \ldots, p$

$$K_i = \frac{R_i - \sum_{k=1}^{i-1} a_k^{(i-1)} R_{i-k}}{E_{i-1}} \tag{25}$$

$$a_k^{(i)} = a_k^{(i-1)} - K_i a_{i-k}^{(i-1)} \qquad \text{for} \qquad 1 \le k \le i-1 \tag{26}$$

$$a_i^{(i)} = K_i \tag{27}$$

$$E_i = \left(1 - K_i^2\right) E_{i-1} \tag{28}$$

The solution for the order p prediction coefficients is then given by

$$a_k = a_k^{(p)} \quad \text{for} \quad 1 \le k \le p \tag{29}$$

In addition to its great computational advantage, the Levinson–Durbin method provides procedural advantages as well. It is important to note that during the recursion steps, the prediction coefficients for predictors of order less than p are calculated, namely, $a_k^{(i)}$ in Eqs. (25)–(27) refers to the kth coefficient of the optimal predictor of order i. Moreover, the prediction error for the lower-order predictors is also produced, $\mathbf{E}_i$ in Eq. (28). This information may be used to select the most appropriate prediction order, p. In contrast, using the standard methods, one would have to solve the equations multiple times to compare the performance of the various order predictors and select the appropriate p. The Levinson–Durbin method also produces an alternate set of coefficients $K_i, 1 \le i \le p$. It is often these coefficients (or a function of them) that are encoded and transmitted. The decoder then uses Eqs. (26) and (27) to reconstruct the prediction coefficients and synthesize the original signal.

One is always faced with the question of which method to choose for estimating the LP coefficients. While there is no rule of thumb, an understanding of the advantages and disadvantages of each may help the reader choose the method most appropriate for the application at hand. For

the autocorrelation method, the advantages are twofold: (1) it utilizes a fast computational algorithm with useful intermediate information and (2) the resulting $A(z)$ is guaranteed to have its roots inside the unit circle, which makes it invertible. This is of great importance since the synthesis filter $H(z)$ becomes unstable if $A(z)$ has roots outside the unit circle. The main disadvantage of the autocorrelation method is the effect of the windowing function on the estimated LP coefficients. The inaccuracies introduced due to windowing lead to suboptimal predictors.

4. EXAMPLE OF LPC

To demonstrate the effectiveness of LPC in signal compression and redundancy removal, consider the speech signal shown in Fig. 3 (this is a portion of the vowel /ee/ as in beet). The signal is sampled at 11,025 samples per second and is 350 samples long. When encoded using a 2-bit per sample pulse code modulation (PCM) encoder, the resulting signal at the receiver is shown in Fig. 4. The coding error, which is the difference between the original and the reconstructed signal, is shown in Fig. 5. The signal-to-noise ratio (SNR) defined as 10 times the base_{10} logarithm of the ratio of the power in the original signal over the power in the coding error is 7.4 dB. If we perform LPC on the original signal using a 10^{th} order predictor, we can then quantize the excitation signal using 2 bits per sample of PCM and use it to synthesize the speech at the decoder (the quantization is performed using adaptive predictive LPC to maximize the SNR [2]). The quantized excitation signal is shown in Fig. 6, while the reconstructed speech signal is shown in Fig. 7. In this case the coding error is shown in Fig. 8. Note that the coding error using LPC is much smaller than the error using PCM. In fact, the SNR for LPC is 19.7 dB, representing a gain of 12 dB at the expense of encoding and transmitting the prediction coefficients and gain (usually on the order of 50 bits). If we use a single bit per sample to encode the excitation signal, thereby cutting the LPC bit rate in half, the resulting signal has a SNR of 13.6 dB, which is still an improvement over 2-bit PCM. In this case, LPC reduced the bit rate and improved the quality of the resulting signal compared to PCM.

5. APPLICATIONS OF LPC

A common model for speech generation (or synthesis) is shown in Fig. 9, where the output speech is produced by passing an excitation signal through an all-pole filter [1,2]. The system shown in Fig. 9 is similar to the synthesis (decoder) system in Fig. 2; the important difference is that the excitation signal is generated locally at the decoder. The nature of the excitation depends on the desired sound to be produced. For voiced speech (e.g., vowel sounds), the excitation consists of periodic pulses. The period of these pulses (also known as the *pitch period*) corresponds to the desired fundamental frequency (or pitch frequency) of the produced sound which varies from speaker to speaker. For unvoiced speech (e.g., fricative sounds such as /s/ and /sh/), the excitation consists of the output of a random-noise generator. Note that, under this synthesis model, the decoder does not need the actual excitation signal. Instead, it needs the type of excitation, the pitch period in case of voiced speech, along with the gain and prediction parameters. Consequently the analysis (encoder) system will differ from the one shown in Fig. 1 and is shown in Fig. 10.

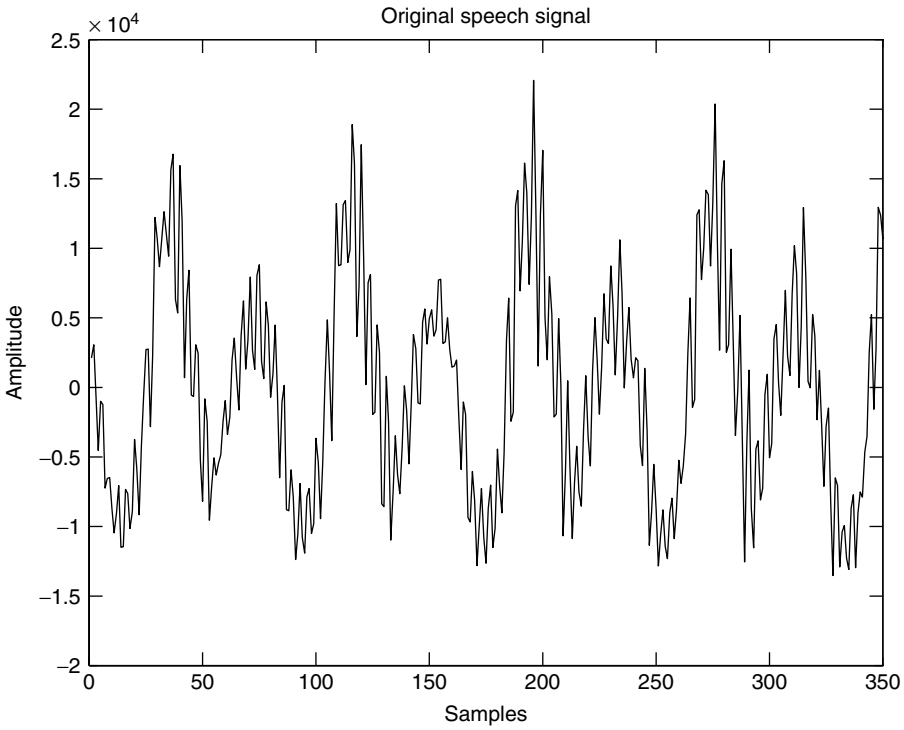


Figure 3. Original speech signal.

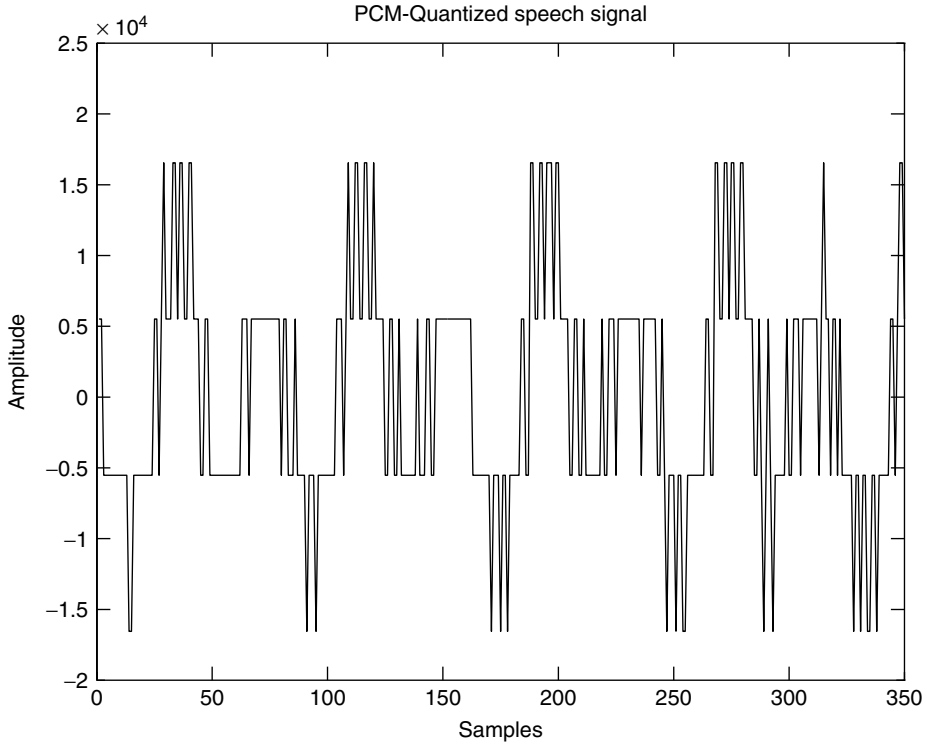


Figure 4. PCM-quantized speech signal using 2 bits per sample.

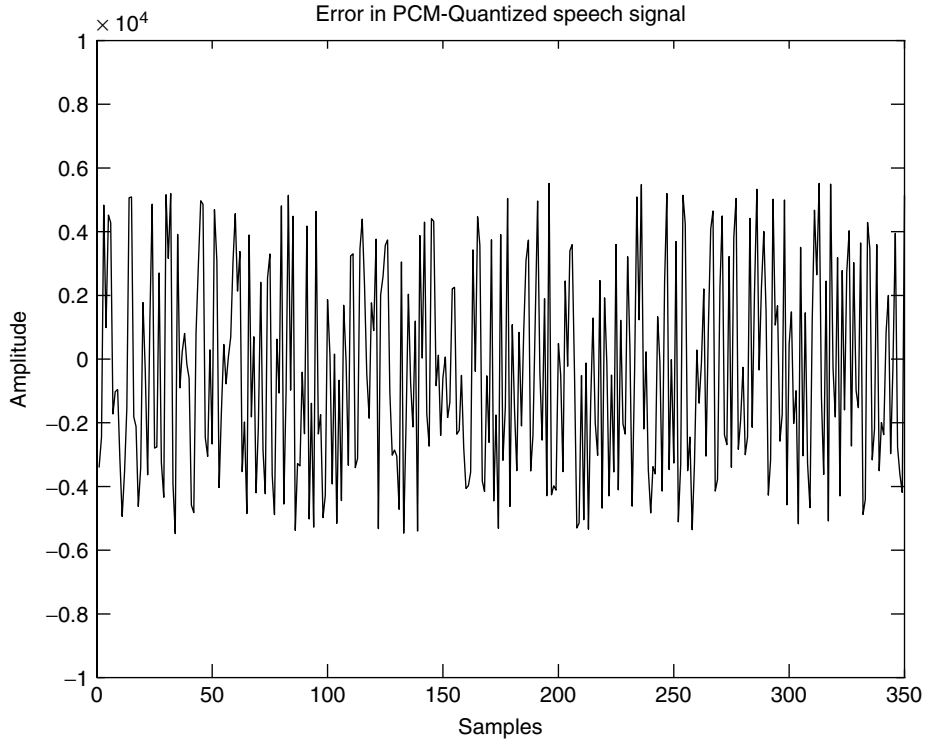


Figure 5. Coding error in PCM-quantized speech signal.

The voiced/unvoiced and pitch information is typically transmitted every 10 ms to track changes in speech. The prediction parameters and gain may be transmitted every 20–30 ms. For speech sampled at 8000 samples per second, it is typical to use a predictor of order 10. Typical bit assignment for the various parameters which would lead to a bit rate on the order of 2400 bits per second (bps) (i.e., approximately 0.25 bits per sample of the original signal). Additional compression may be obtained by applying vector quantization to the parameters. If we had used PCM to encode the original speech, the required bit rate would be on the order of 64,000 bps.

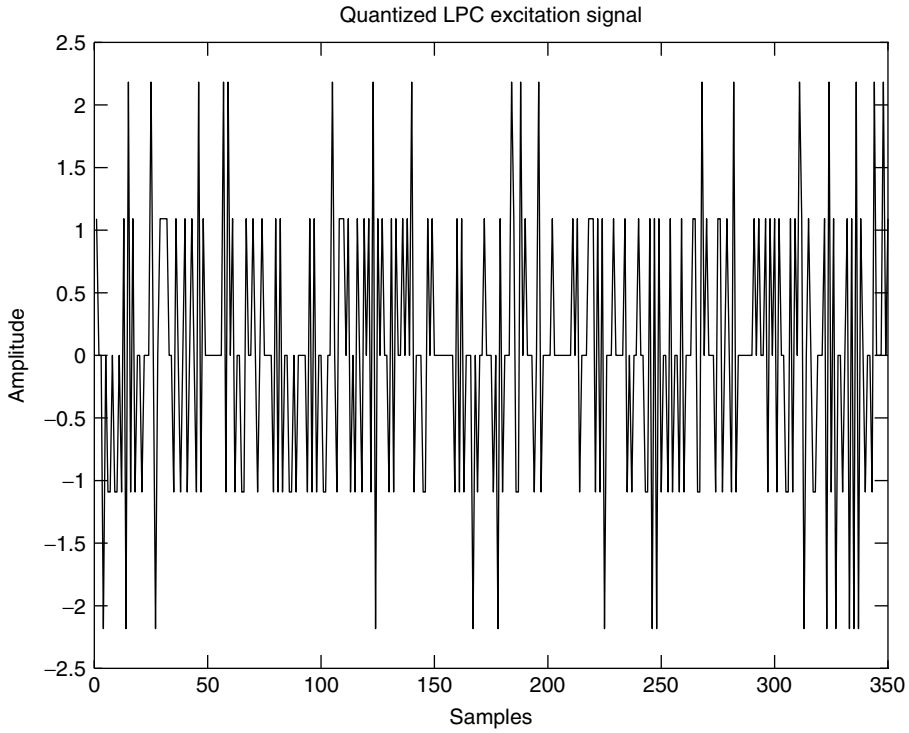


Figure 6. Quantized LPC excitation signal.

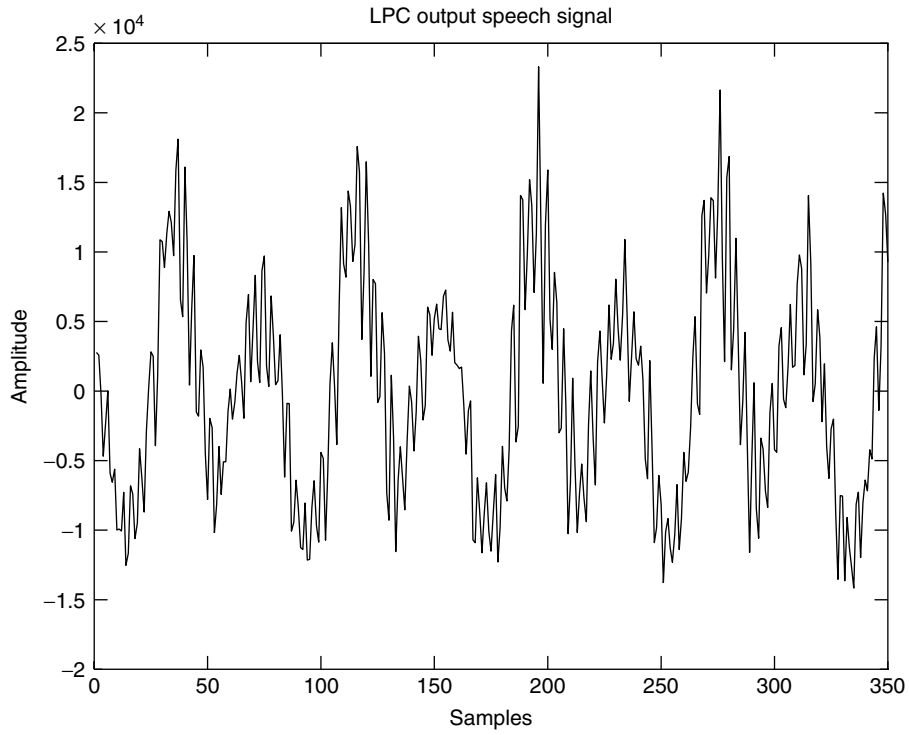


Figure 7. Reconstructed speech signal using LPC.

Clearly, LPC provided great reduction in bit rate in this case. Some of this reduction, however, comes at the expense of the quality of the output speech. Because of the synthesis model used with the hard switching between sources of excitation, the resulting speech lacks naturalness and is often described as choppy. Over the years, more sophisticated variations on LPC have been devised. We describe two of the more recent methods below: code-excited LP (CELP) [5] and mixed-excitation LP (MELP) [6].

In CELP, the excitation signal is selected from a library (referred to as a *codebook*) of possible excitation signals. The encoder conducts an exhaustive search of all the excitations in the codebook to determine the one that

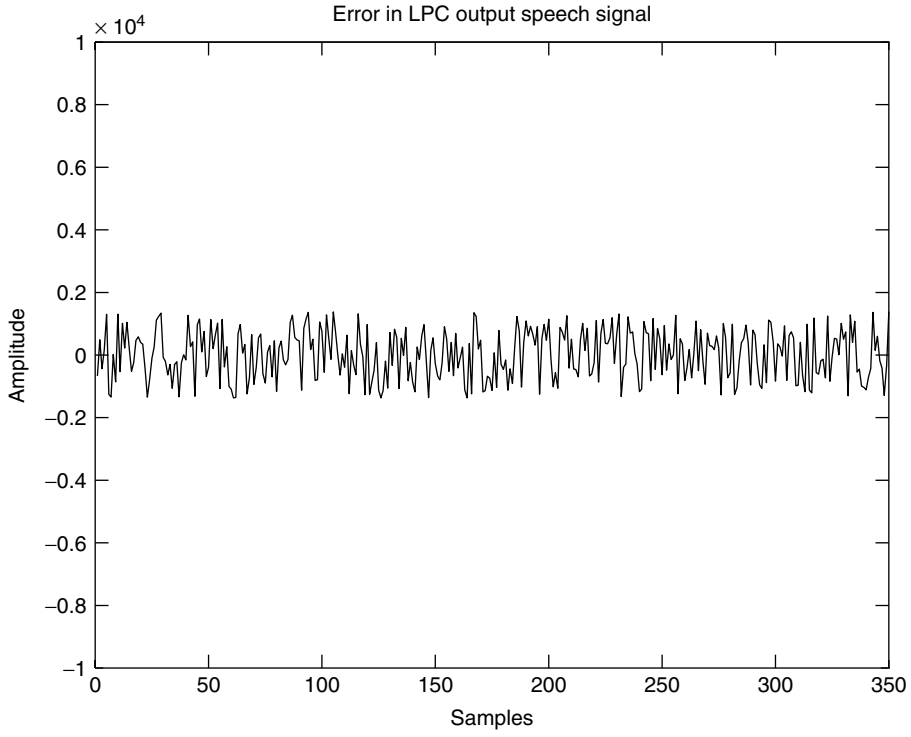


Figure 8. Coding error in LPC generated speech signal (same scale as in Fig. 5).

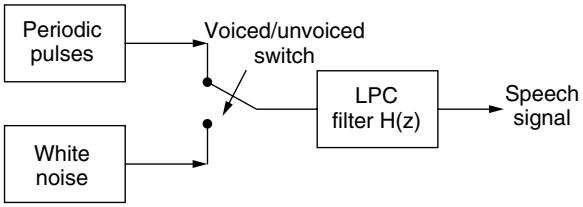


Figure 9. Typical speech synthesis system using LPC.

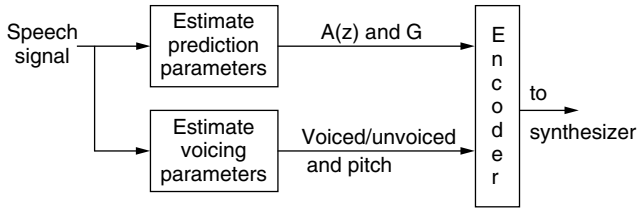


Figure 10. Speech analysis system for use with the synthesizer in Fig. 9.

produces the best output speech, then transmits the index of this best excitation to the decoder along with the LP parameters and gain. If the codebook is composed of 1000 possible excitations, for example, 10 bits would be needed to encode the index information. This information is transmitted every 5 ms; consequently 2000 bps are required for the excitation information. Often in CELP, two predictors are employed. The first predictor is used to remove the short-term redundancy as described earlier, while the second predictor is used to remove the long-term redundancy associated with the periodicity of voiced speech. In the latter case, the predicted sample is no longer based on neighboring samples, but on ones a pitch period away. In essence the long-term predictor carries the pitch information. The set of parameters associated with CELP are therefore the excitation index in the codebook, the LP coefficients, the long-term predictor coefficients, and the delay of the long-term predictor. CELP offers improved output speech quality when compared to the system in Figs. 9 and 10 at the expense of an increase in bit rate to around 4800 bps.

In MELP, the excitation signal is chosen as a combination of the excitation functions in Fig. 9. Each excitation function (i.e., periodic pulses and white noise) is passed through a multiband filter before they are combined. As a result, the excitation signal is considered to be voiced in some frequency bands and unvoiced in others. This model reflects more closely the true nature of speech and, as a result, produces more natural output speech. The encoder in MELP transmits the voiced/unvoiced information associated with each band, as well as pitch information, gain and LP coefficients. Using efficient encoding and quantization, MELP produces good quality speech, rivaling that of CELP at about half the bit rate, namely, 2400 bps. Methods aimed at reducing the bit rate are currently under investigation with goals of LPC-based coders operating in 600–1200 bps range while producing high quality natural-sounding speech.

BIOGRAPHY

Amro El-Jaroudi was born in Cairo, Egypt, in 1963. He received his B.S. and M.S. degrees in 1984, and his Ph.D. in 1988 from Northeastern University in electrical engineering. In 1988, he joined the Department of Electrical Engineering at the University of Pittsburgh,

Pennsylvania, where he is now associate professor. His research interests include digital processing of speech signals with applications to speech coding and recognition, spectral estimation of nonstationary signals, and pattern classification.

BIBLIOGRAPHY

1. J. Makhoul, Linear prediction: A tutorial review, *Proc. IEEE* **63**(4): 561–580 (April 1975).
2. J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, New York, 2001.
3. C. Lee, Robust linear prediction for speech analysis, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, April 1987, pp. 289–292.
4. A. El-Jaroudi and J. Makhoul, Discrete all-pole modeling, *IEEE Trans. Signal Process.* **39**(2): 411–423 (Feb. 1991).
5. M. R. Schroeder and B. S. Atal, Code-excited linear prediction (CELP): High quality speech at very low bit rates, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, March 1985, pp. 937–940.
6. A. V. McCree and T. P. Barnwell III, Mixed excitation LPC vocoder model for low bit rate speech coding, *IEEE Trans. Speech Audio Process.* **3**: 242–250 (July 1995).

LOCAL MULTIPOINT DISTRIBUTION SERVICES (LMDS)

Peter Papazian
Roger Dalke
Institute for Telecommunication Sciences
Boulder, Colorado

1. INTRODUCTION

LMDS is the acronym for Local Multipoint Distribution Service, a broadband wireless access (BWA) service being developed in the United States at millimeter wave(length) frequencies. Similar BWA services, sometimes with different names but also at millimeter wave frequencies, are concurrently being developed and deployed in Canada [Local Multipoint Communication Service or (LMCS)], Europe, Asia, and Central and South America. As the name implies, LMDS is a short-range (local), point-to-multipoint broadcast service. The service will allow two-way communication and has been allocated more than 1 GHz of radio spectrum in the United States. This large bandwidth enables high-speed (high-bit-rate) wireless communication. LMDS is envisioned as a wireless link to a metropolitan-area network (MAN) capable of providing simultaneous interactive video, digital telephony, data, and Internet services. These services are allowed two modes of operation: point-to-point and broadcast. The point-to-point mode operation is similar to fixed microwave links. However, larger link budgets must be allocated for signal fading due to rain and for attenuation due to atmospheric adsorption. Point-to-point radio links can serve medium to large size business customers and have also been used to provide service to niche markets, small areas not served by cable or urban buildings where cable or fiber would be too expensive to install. The broadcast service was initially envisioned as providing internet, video, and telephony services to consumers on a large scale. This market has been slow to develop due to the costs of infrastructure development and the technical difficulties of obtaining adequate signal coverage. Both of these factors have made these systems economically unfeasible to deploy in the United States so far.

The advantages and disadvantages of LMDS are related to the use of the extremely high/superhigh frequency (EHF/SHF) or millimeter wave portion of the radio spectrum. The millimeter wave spectrum allows some equipment miniaturization and has large available bandwidths necessary for high-speed digital communication. But the high radiofrequencies also cause problems due to radiowave propagation impairments, the higher cost of electronic components and unavailability of high power solid-state linear amplifiers. For a summary of the most recent advances in amplifier technology, see Ref. 1.

The remainder of this article is organized in the following manner. First an overview of the LMDS band (spectrum) allocation and some technical rules related to the use of this band are given. This section also discusses work done by standards groups to help speed development and deployment of LMDS. Then millimeter wave radiowave propagation impairments are presented analytically. Finally, radiowave propagation measurements for an LMDS broadcast system are summarized.

2. REGULATORY AND STANDARDS OVERVIEW

Figure 1 is the band allocation chart for LMDS in the United States as specified by the Federal Communications Commission (FCC) [2]. Under this band plan, two blocks of frequencies (A and B) near 30 GHz are allocated in 493 basic trading areas (BTAs). These areas are defined in the *Rand McNally Commercial Atlas* and details about BTAs can also be obtained from the FCC website (*www.fcc.gov*). The LMDS spectrum was then licensed by block and BTA to successful bidders at the FCC LMDS spectrum auction. Block A has 1150 MHz of radio spectrum, and block B has a 150-MHz allocation. A critical issue for LMDS operation is the task of avoiding interference at BTA boundaries and in shared bands. This task is called *frequency coordination and interference control*. Frequency coordination and interference rules are more difficult to define for systems that can operate in the broadcast mode such as LMDS than for point-to-point operation more typical in the microwave bands. Usually the procedure requires a knowledge of the following transmitter and receiver parameters: (1) *effective isotropic radiated power* (EIRP) or *power flux density* (PFD), (2) channelization and frequency plan, (3) modulation type and channel bandwidth, (4) frequency stability, (5) receiver parameters (noise figure, bandwidth, and

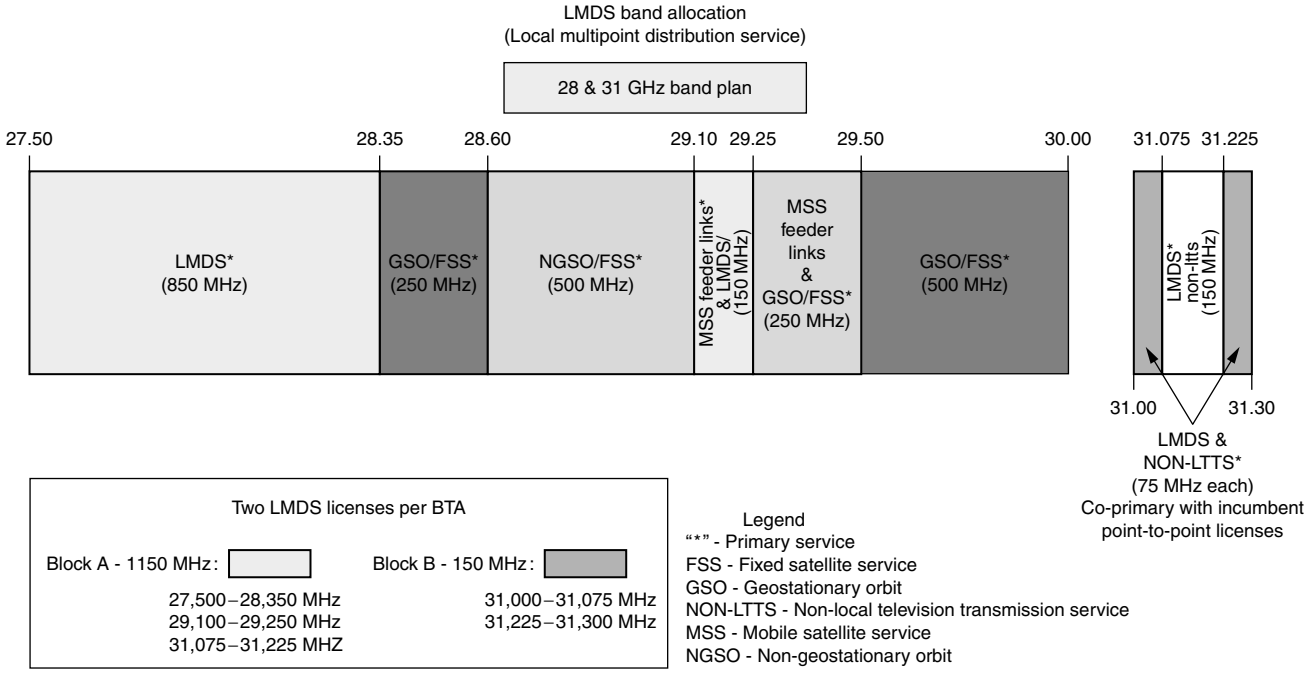


Figure 1. LMDS band allocation chart for the United States, source FCC.

thresholds), (6) antenna characteristics, and (7) system geometry. The FCC declined to set these standards, citing the technical difficulties of calculating a reasonable limit and a lack of support by industry for establishment of such a limit. Instead, frequency coordination between adjacent BTAs was left to a cooperative effort between license holders as specified in Section 101.103(d) of the *Code of Federal Regulations* (CFR) [3]. These coordination rules are applied to LMDS stations within 20 km of the BTA boundary. Within each BTA and frequency block, operators are left with the task of establishing their own frequency coordination rules to avoid interfering with adjacent hubs in their own cellular-type broadcast or point-to-point system. The FCC did set maximum allowable EIRP for any system by frequency band; these limits are listed in Table 1.

Referring to Fig. 1, we see that between 29.1 and 29.25 GHz, LMDS coexists with mobile satellite feeder links. These coexistence rules are specified by the FCC to protect existing satellite links. Since LMDS stations can transmit in the point-to-point mode as well as in the broadcast mode, two types of EIRP coexistence rules were specified. For point-to-point narrowband operation, the EIRP per carrier is limited to −23 to −26 dBW/MHz, depending on climate zone. To prevent LMDS broadcasting base stations from interfering with mobile satellite stations, the EIRP aggregate power spectral density per unit area for all LMDS hub transmitters in a BTA is limited to between −23 and −26 dBW/MHz · km^2 [3,4].

Table 1. FCC EIRP Limitations by Band for LMDS Systems

	Maximum Allowable EIRP	
Frequency Band (GHz)	Fixed (dBW/MHz)	Mobile (dBW/MHz)
27.50–28.35	30[a]	—
29.10–29.25	−23 to −26[b]	—
31.00–31.075	30	30
31.225–31.30	30	30

[a] 42 dBW/MHz for subscriber terminals.
[b] See text.

It was envisioned by the FCC that each operator would install a sufficient number of base stations in the BTA to meet subscriber demand, develop interference and coexistence rules, and manage frequency reuse in their own frequency block. The IEEE Wireless LAN/MAN Standards Committee group, 802.16, has been convened to develop these coexistence and channelization rules. This will be accomplished by defining the Physical and media access control (MAC) layer standards for proposed LMDS systems. At present only a draft standard is available. The work of the 802.16 group can be retrieved online from the IEEE web site (*www.ieee.org*). Results from 802.16 that are available indicate some of the channelization and the capacity or spectral efficiency of the proposed LMDS and BWA systems both in the United States and abroad. See Table 2 for a summary of these proposed frequencies and aggregate transmission rates. It should be noted that 802.16 regards its work as encompassing both LMDS and other BWA systems in the millimeter wave and microwave frequency range. This can cause some difficulty when trying to focus on LMDS standards.

The FCC has also required operators to provide a substantial level of service in their BTA. For an LMDS license that provides point-to-multipoint service, coverage to 20 percent of the population in the service area at the 10-year mark would constitute substantial service. For a license holder choosing to deploy point-to-point service,

Table 2. Summary by Country of Frequency Allocations and Estimated Aggregate Data Rates for LMDS and BWA Services Operating in the Millimeter-Wave Frequency Range

Country	Frequency (GHz)	Bandwidth (MHz)	Proposed Rates[a] (Mbps)
USA	LMDS block A 28,29,31	1150	862
	LMDS block B 31	150	115
	38 (point to point)	$N \times 50$	$N \times 75$
	40	Future	
	60	Future	
Canada	LMCS 25–28	3000	
Japan	23–28	Various	
Europe	26	Various	
	40	3000	
	28	LMDS equivalent	
Korea	25–27		
Asia	26–31,38	Various	

[a]Total data rate for the entire frequency block.

four permanent links per million people in the service area at the 10-year mark would constitute substantial service. More details on these rulings can be found in Refs. 2–4.

3. MILLIMETER WAVE PROPAGATION

3.1. Clear Air Absorption

At frequencies above 10 GHz, radiowaves propagating through the atmosphere are subject to molecular absorption. Although typical LMDS frequencies near 28 GHz are in a "window" — comfortably between the water vapor absorption line at 22 GHz and the band of oxygen lines near 60 GHz — there will nevertheless be some residual effects from the tails of these and other lines. Such effects can be evaluated using the millimeterwave propagation model of Liebe [5,6] (see also Rec. ITU-R P.676-4 [7]). For example, Fig. 2 shows clear air absorption as a function of frequency and humidity for a standard atmosphere (15 °C, 1013.25 mbar) and a frequency range of 10–100 GHz. Figure 3 shows the absorption as a function of relative humidity and temperature for 28 GHz and 1013.25 mbar. Note that on a hot, muggy day, a 6-km path could suffer perhaps 5 dB clear air attenuation.

3.2. Effects of Rain

Absorption and scattering of radiowave energy, due to the presence of raindrops, can severely degrade the reliability and performance of communication links. Attenuation resulting from propagation through raindrops is perhaps the most significant threat to line-of-sight (LoS) radio links operating in the millimeter waveband. For such systems, reliability predictions based on rain attenuation alone are often sufficient, since the error due to the exclusion of the other atmospheric propagation effects is much less than

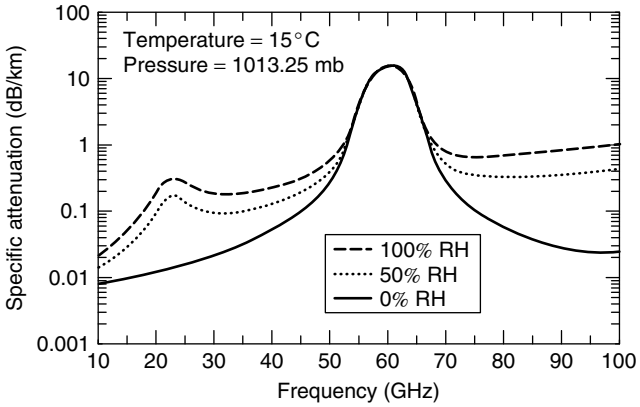


Figure 2. Clear air absorption as a function of frequency and humidity at 15 °C and 1013.25 mbar.

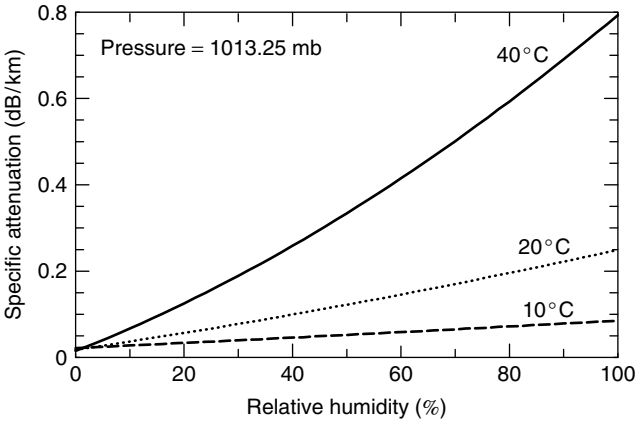


Figure 3. Clear air absorption as a function of relative humidity and at 28 GHz and 1013.25 mbar.

the normal year-to-year variation in rain attenuation. In general, rain-induced dispersion (frequency selectivity) is not considered significant for bandwidths of less than 1 GHz [8]. Note that proposed channelization schemes within the LMDS bands are all much smaller than 1 GHz, hence this effect can be ignored.

Rain attenuation is a function of drop shape, drop size, rain rate, and wavelength. Since the drops are randomly distributed in the atmosphere, the net scattering is an incoherent superposition of contributions from individual drops. The power scattered and absorbed per drop for a unit incident energy flux is called the *absorption cross section* σ, which for spherical drops is a function of the wavelength, drop radius, and the refractive index.

In traversing an incremental distance ds through spherical raindrops of radius r, the fractional loss of flux is $n_r \sigma ds$, where n_r is the number of drops per unit volume with radius r. The beam intensity decays exponentially, i.e., $I(x) = I_0 e^{-ax}$, where $a = n_r \sigma$ is the attenuation or extinction coefficient. In a rain storm, the actual drop sizes vary with rain rate and type of storm activity, and hence, the total attenuation is obtained by summing the

contribution from all drop sizes or

$$a = \int \sigma(r, \lambda, m) n(r)\, dr$$

where $n(r)$ is the drop size distribution, λ is the wavelength, and m is the complex refractive index. The specific attenuation over a path of length L in decibels per unit length is

$$\alpha = \frac{10 \log_{10}\{I_0/I(L)\}}{L} = 4.343\, a$$

The drop size distribution is a function of rain rate and type of storm activity and is well represented by an exponential of the form [9]

$$n(r) = N_0 e^{-cR^{-d}r}$$

where R is the rain rate, in mm/hr, r is the drop radius in millimeters, and c and d are empirical constants. The absorption cross section can be calculated using the classic scattering theory of Mie for a plane wave incident on an absorbing sphere [9]. For frequencies of $\leq$40 GHz, where the wavelength is much greater than the drop size, the *Rayleigh approximation* can be used. The Rayleigh scattering cross section is given by

$$\sigma = \frac{8\pi^2}{\lambda} r^3 \mathrm{Im}\left[\frac{m^2 - 1}{m^2 + 2}\right]$$

where Im refers to the imaginary part of the argument.

Integrating over all possible drop sizes and assuming Rayleigh scattering gives a relatively simple relationship between the specific attenuation and the rain rate: $\alpha = aR^b$ (dB/km). The coefficients a and b depend on the drop size distribution, refractive index, and frequency. By convention, the coefficients are given for rain rates in mm/h. This result is consistent with direct measurements of attenuation and is in agreement with Mie scattering [9] over a wide frequency range.

Several investigators have studied the distribution of raindrop sizes as a function of rain rate and type of storm activity. Olsen et al. [10] give tables of coefficients for several spherical drop size distributions as a function of temperature and frequency. The most commonly used distributions are those of Law and Parsons (LP), Marshall and Palmer (MP), and Joss and Waldvogel (JW). Law and Parsons propose two distributions, the LP(L) distribution for widespread rain (with rates less than 25 mm/hr), and the LP(H) distribution for convective rain with higher rates. In general the LP distributions seem to be favored for design purposes because they have been widely tested and compared to measurements. The LP(L) distribution gives approximately the same specific attenuation as the JW thunderstorm distribution, and the specific attenuation of the MP and LP(H) are approximately the same. Allen [11] points out that for millimeter waves there is a range of more than a factor of 2 in specific attenuation for different drop size distributions used by Olsen et al. and a range of a factor of 4 for the different climate regions used by Dutton et al. [12]. The resulting uncertainty is a critical limitation to predicting link reliability.

In general, drops are not spherical, in which case the coefficients depend on polarization. Coefficients for vertically and horizontally polarized electromagnetic waves have been calculated for oblate spheroidal drops using methods similar to those described above. Coefficients for nonspherical drops and methods for calculating coefficients for arbitrary polarizations are given in ITU-R P.838-1 [13]. Figure 4 compares the specific attenuation at 30 GHz as a function of rain rate for L-PL coefficients (spherical drops at 20 °C) and ITU coefficients for horizontal and vertical polarization. Note that vertical polarization provides a significant advantage for lengthy paths in moderate to severe rain.

In principle, the total attenuation is obtained by integrating the specific attenuation over a particular path. Accurately modeling the total attenuation is difficult since rain rate is a nonstationary random process with short and long term, as well as global and local variations. Local variations occur because the vertical distribution of precipitation varies with temperature as a function of height. Also, intense rain tends to be localized and the rain rate can vary significantly over terrestrial paths.

Two important models that are commonly used to predict terrestrial path attenuation are the global model of Crane [14] and the ITU model (ITU-R P.530-8 [15]). Both models provide empirical formulas for calculating an *effective pathlength* L_{eff} that is a function of the rain rate. The path attenuation is then the product of the specific attenuation based on the locality or *point* rain rate and the effective pathlength

$$A(\mathrm{dB}) = aR^b L_{\mathrm{eff}}$$

If measured rain rate statistics for the desired location are not available, both the Crane global model and the ITU model give methodologies for estimating rain rate statistics for an *average year*. The ITU model provides global data for calculating rainfall statistics with grid points spaced at 1.5° intervals in both latitude and longitude. The Crane global model partitions the world into 12 rain climate zones based on the assumption that the location-to-location variability within a zone

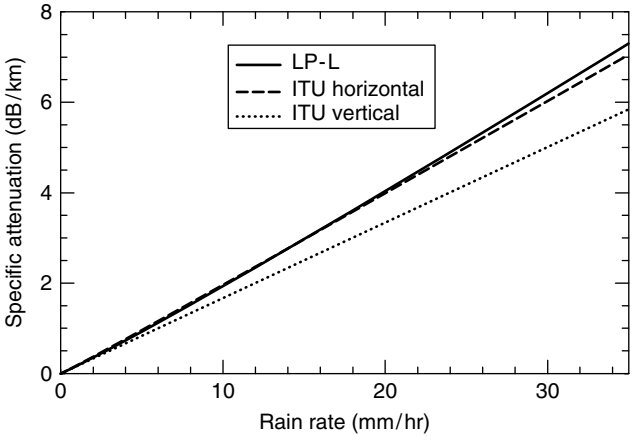


Figure 4. Specific attenuation as a function of rain rate for LP(L) (spherical drops at 20 °C) and ITU drop distributions.

is comparable to the year-to-year variation at a point. Location-to-location variability and year-to-year variation in the rain rate distribution are reported to be lognormal with a standard deviation of 50% for all climate regions [16]. The Crane model is widely used and has been shown by Dutton [17] to be one of the better models.

The Rice–Holmberg [18] global surface rain rate model can be used to calculate local rain rate statistics using historical meteorological data. This model is based on extensive long term rain rate statistics from 150 locations throughout the world. The Rice–Holmberg model gives the rain rate distribution in terms of commonly recorded climatologic parameters: the average annual rainfall accumulation and the average annual accumulation of thunderstorm rain. According to this model, the cumulative distribution of 1-minute average rainfall rates for an *average year* is given by

$$P\{R > \rho\} = \frac{M}{8766}\{0.03\beta e^{-0.03\rho} + 0.2(1-\beta) \times [e^{-0.258\rho} + 1.86e^{-1.63\rho}]\}$$

where M is the average annual rainfall accumulation in mm and β is the average annual ratio of thunderstorm rain to total rain. The required climatological data can be obtained from a variety of sources. Perhaps the best source for the rainfall data is the National Climatic Data Center (NOAA/National Weather Service, Asheville, North Carolina) which is the world's largest active archive of weather data. Extensions to the Rice–Holmberg model that include year-to-year and location-to-location variability are given by Dutton [19].

As an example, consider radio links of less than 20 km in the vicinity of San Francisco, California. Using the Rice–Holmberg model and historical data from a local weather station [20], the 1-min rain rate exceeded less than .01% of an average year is found to be 18.6 mm/h. Considering the LP(L) and the ITU coefficients for vertically and horizontally polarized waves, the specific attenuation for 18.6 mm/h is obtained from Fig. 4. The effective path lengths calculated using both Crane global model and ITU procedures are shown in Fig. 5. The total path attenuation is the product of the specific attenuation and the effective pathlength that corresponds to the actual pathlength. For a 10-km link assuming an LP(L) drop distribution ($\alpha = 3.73$ dB/km), the total path attenuation is 37.3 dB according to the Crane global model and 27.1 dB according to the ITU model. These results give the total attenuation exceeded less than 0.01% of an average year.

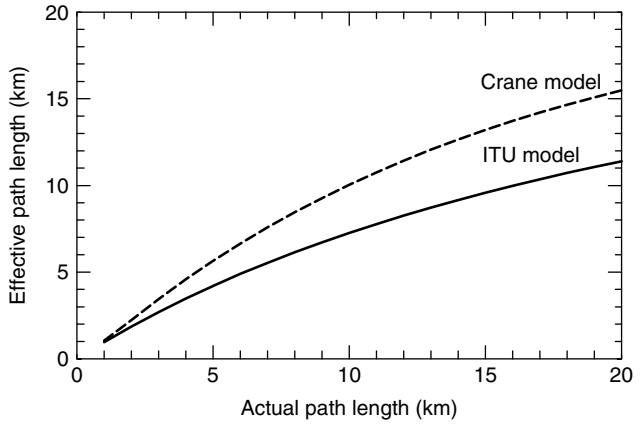


Figure 5. Effective pathlength based on the Crane global and ITU models assuming a rain rate of 18.6 mm/h.

3.3. Rain-Induced Depolarization

Rain-induced depolarization is due to differential attenuation and phase shifts caused by nonspherical raindrops. The classic model for a falling raindrop is an oblate spheroid with its major axis canted to the horizontal and with major and minor axes related to the radius of a sphere of equal volume. For practical applications a semiempirical relationship between rain attenuation and depolarization is provided by Ippolito [9] (see also Ref. 15):

$$\mathrm{XPD} = 30\log_{10} f_{\mathrm{GHz}} - 10\log_{10}(0.5 - 0.4697\cos 4\tau) - 40\log_{10}(\cos\theta) - 23\log_{10} A$$

where XPD is the "cross-polarization discrimination," that is, the ratio (in decibels) of the copolarized and cross-polarized field strengths, where τ is the tilt angle of the polarization with respect to horizontal, θ is the elevation angle of the path, and A is the rain attenuation in decibels. For 30-GHz terrestrial links ($\theta \approx 0$) with attenuation of less than 15 dB, and horizontal ($\tau = 0$) or vertical ($\tau = \pi/2$) polarization, the effects of rain-induced depolarization are quite small (XPD > 30 dB).

3.4. Attenuation Due to Fog

Fog results from the condensation of atmospheric water vapor into water droplets that remain suspended in air. There are two main types of fog. Advection fog is coastal fog that forms when warm, moist air moves over colder water. Liquid water content of advection fog does not normally exceed 0.4 g/m^3. Radiation fog forms inland at night, usually in valleys and low marshes, and along rivers. Radiation fog can have a liquid content of up to 1 g/m^3.

Specific attenuation for fog can be calculated using a model developed by Liebe [6]. Using this model and assuming dense fog conditions with 1 g/m^3 water result gives a specific attenuation of 0.5 dB/km. For a homogeneous fog path of 6 km, the total attenuation is 3 dB.

4. LMDS RADIO CHANNEL

Signal impairments due to atmospheric gases, rain, and rain depolarization are important radiowave propagation factors that can be computed using methods outlined in the previous sections. However, millimeter wave signal dispersion due to multipath and attenuation, and depolarization due to random distributions of vegetation are system and site dependent and must be measured. For example, multipath measurements are highly dependent on the beamwidth of the transmitting and receiving

antennas. Since narrow beam antennas will filter out multipath signals, a multipath metric such as delay spread (S), will be smaller for LMDS point-to-point systems using a narrow beamwidth antenna than an LMDS broadcast system using a wider beamwidth antenna. The site dependence of these parameters for broadcast systems is also critical. For instance, the percentage of LoS paths to potential subscribers (and hence signal attenuation) will vary depending on transmitter location and environment. An urban high-rise or hilly suburban environment will suffer more blocked paths than a flat rural environment with little vegetation. When dealing with attenuation due to vegetation, further environmental distinctions must be made to account for vegetation type, density, and distribution.

Fortunately some aspects of the millimeter wave radio channel can be modeled using advanced computer methods. Standard radio propagation models can calculate diffraction and signal blockage due to terrain by incorporation of digital terrain data for an area. More advanced models have been developed for millimeter wave and LMDS applications that incorporate higher resolution terrain and building elevation data from aerial photographs. These programs can then determine LoS

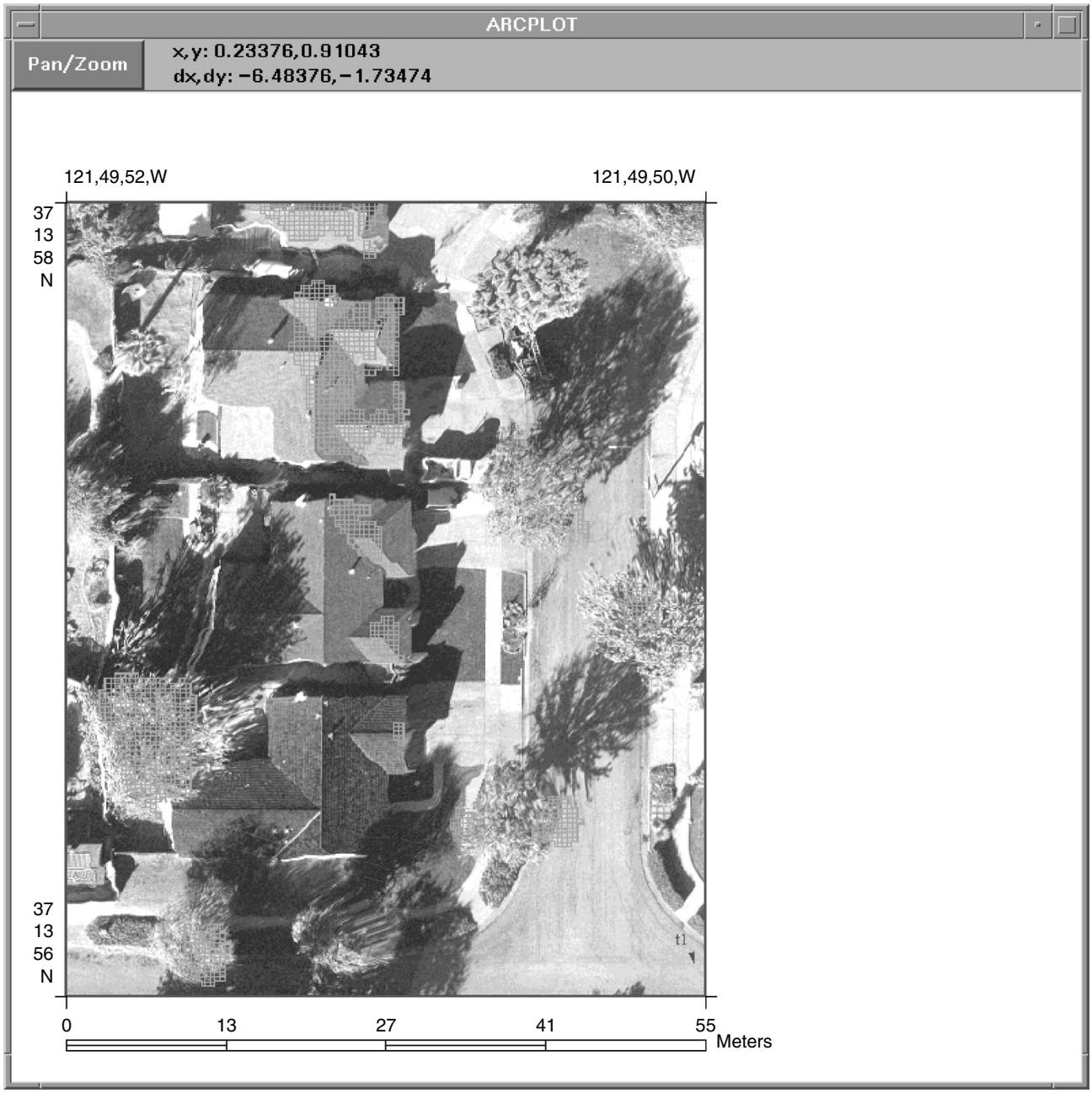


Figure 6. LMDS propagation modeling program output overlaid on an aerial photograph of San Jose, CA. Purple areas indicate line-of-sight coverage from a 12 m high transmitter located to the southeast.

Table 3. Measurement Equipment Parameters

	Antenna Beam Width (degrees)		EIRP (dBm)	Sensitivity (dBm)	
	Vertical	Azimuthal		Narrowband	Wideband
Transmitter	20	90	51	N/A	N/A
Receiver	7.5	7.5	N/A	−130	−102

paths at millimeter wave frequencies. Figure 6 is an example of such a computer simulation.[1] This figure is an aerial photograph of several houses in San Jose, CA. The purple areas indicate LoS coverage from a 12-m transmitter located to the southeast. Computer processing of the photograph was used to develop the surface contour used in conjunction with latitude, longitude and elevation of the transmitter location to determine LoS paths. However, these efforts still lack the ability to incorporate multipath and attenuation, diffraction, and depolarization due to vegetation. To incorporate these effects, measurement data are required.

4.1. LMDS Broadcast System Area Coverage and Radio Channel Measurements

LMDS is ideal for providing last-mile connectivity to a fixed, broadband network. To achieve this it is important to know the area coverage and the radio channel characteristics for specific sites and proposed systems, including vegetation. A typical last mile solution in suburban neighborhoods will utilize broadcasting base stations arranged in a cellular pattern with low antenna heights and spaced on a 1–2-km grid. Consumers could then install small directional antennas aimed at the base station. Typically the forward link from the base station would be high power and high bit rate while the reverse link would be low power and low bit rate. To quantify the percentage of households that can be reached (coverage achieved), as well as the radio channel characteristics, a set of 30-GHz radiowave propagation measurements is described below.

These measurements include area coverage, signal attenuation, signal depolarization, and the delay spread (S) for an LMDS radio channel. The measurement system transmits a 28.8-GHz narrowband continuous-wave (CW) signal and a 30.3-GHz wideband signal through a common traveling wavetube amplifier. The wideband signal, used to measure the radio channel impulse response, was created by modulating the carrier with a 500-Mb/s pseudo-random-noise code. The transmitter used a single vertically polarized horn with 14 dB gain. The antenna had a 90° azimuthal 3-dB beamwidth, and a 20° vertical beamwidth. The EIRP for the transmitter was 51 dBm (−6 dBW/MHz for 500 MHz BW).

The receiver antenna system consisted of two 7.5° dishes with linearly polarized feeds. One dish was aligned for vertical polarization, and the second was aligned for horizontal polarization. The received signals were split and processed in separate narrowband and wideband receivers. The wideband receiver provided cophase and quadrature-phase impulse response data with 2-ns resolution. The receiver had a sensitivity of −102 dBm and a dynamic range of 50 dB. The narrowband receiver was used to measure received signal power. It had a sensitivity of −130 dBm and a dynamic range of 70 dB. Some relevant equipment parameters are listed in Table 3.

4.2. Environment

The measurement area consisted of one- and two-story single-family residences in Northglenn, Colorado and San Jose, California. Both areas have small yearly rainfall totals and slow tree growth. Some relevant geographic statistics for each site are listed in Table 4.

Factors that can affect coverage at these sites include rainfall, terrain, shadowing by buildings, and attenuation by vegetation. Since the terrain at both survey sites is flat, this is not an issue when comparing results between the two sites. The distributions of roof heights for each site were estimated from measured data and are also similar. The most important difference between the sites is the vegetation, in particular the tree canopy. The tree population in Northglenn is dominated by elms, maples, cottonwoods, and ponderosa pines. Mature trees of these species are 9–15 m tall. In contrast, many trees in San Jose have tropical origins and are only 6–9 m tall.

4.3. Measurement Procedures

Both narrowband and wideband data were collected. The narrowband data includes a time series record of the signal power, which was used to study area coverage, short-term variations of the signal and depolarization. These data were recorded at 1000 samples/s for 50 s. Wideband data, used to measure multipath, consisted of 100 complex impulse responses at each site. Each impulse lasted for 254 ns and was sampled 1000 times. The repetition rate of

Table 4. House Density, Normal Temperature, and Rainfall Averages for Northglenn, CO, and San Jose, CA

Geographic Statistics	Northglenn, Colorado	San Jose, California
Number of houses/km^2	780	900
Temperature (°F)[a]	50.3	59.7
Precipitation (in.)[b]	15.31	13.86

[a] Monthly average.
[b] Yearly normal between 1951 and 1980.

[1] The user's guide to CSPT (communications system planning tool) is available from ITS by request. The software is available free of charge to users.

the impulses from the sliding correlator was 10 Hz. Both data sets were collected using vertical (copolarized) and horizontal (cross-polarized) receive antennas.

The receiver address was determined by randomly selecting houses, using aerial photographs of the survey area. Because it was assumed that the probability of acceptable coverage would decrease with distance, each broadcast cell was first subdivided into bands of increasing radii from the transmitter. Stations (houses) were then selected randomly from equal area subdivisions of each band. Figure 7 shows a typical 0.5-km square cell quadrant with its three sampling bands. The number of stations needed for an acceptable error was determined by assuming that the area coverage estimate could be modeled using a binomial distribution (see the next section for a description of this model).

At each receiver station the curbside location of the measurement van was selected using both aerial photographs and onsite inspection to avoid obvious obstructions between the roof of the house and the transmit antenna. The receive antenna height was determined using a mast-mounted videocamera to locate the height of the roof peak above street level and then by raising the mast an additional meter. The optimum receiver antenna azimuth and elevation angle were then determined using narrowband, vertically polarized, azimuth, and elevation scans to find the direction of maximum received power.

It was desired to estimate coverage for cells that could be separated into four symmetric quadrants. To save time, only one quadrant of each cell was sampled. It was assumed that the other quadrants would be sufficiently uniform and would produce similar results for the entire cell. In Northglenn, two 0.5-km (Fig. 7) square cell quadrants were surveyed using different 12-m-high transmitter locations, and the area coverage results were compared and found to yield similar results. In San Jose, one 0.5-km square cell quadrant and a 1-km circular cell quadrant were surveyed using a 12-m-high transmitter site. To study the area coverage dependence on transmitter height, a 24-m-high transmitter site was added. The 0.5-km quadrant and 1-km quadrant were re-surveyed using this transmit antenna. Then a 2-km circular cell quadrant was surveyed, also using the 24-m transmit antenna to study the coverage dependence on transmitter height.

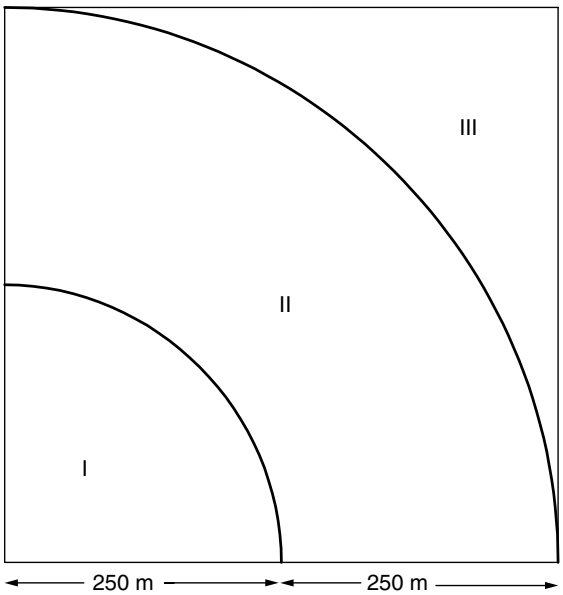


Figure 7. 0.5-km cell square cell quadrant with transmitter located in the lower left corner. Areas I, II, III indicate sampling zones for the calculation of area coverage statistics.

4.4. Area Coverage Model

The area coverage estimates are based on copolarized (vertical) received power data. The coverage in each cell band can be estimated as the fraction of houses for which an adequate signal is available for a given percentage of the time. If p_i is the area coverage probability for the band, n_i is the number of houses sampled, and n_{i1} is the number of houses in the ith band that meet the signal level requirements for coverage, the area coverage estimate is

$$p_i = \frac{n_{i1}}{n_i}$$

Assuming that the number of houses with coverage is binomially distributed and the area is sampled without replacement, the standard error σ_i in each cell band can be approximated as [21]

$$\sigma_i = \sqrt{p_i(1-p_i)\left(\frac{1}{n_i} - \frac{1}{N_i}\right)}$$

where N_i is the number of houses in the ith band. The area coverage p_c and error estimates σ_c for each cell are calculated by weighting the results from each band using their relative area a_i and summing the results from each band as follows:

$$p_c = \sum_{i=1} a_i p_i$$

$$\sigma_c = \sqrt{\sum_i a_i^2 \sigma_i^2}$$

4.5. Area Coverage Metric

The metric used to determine area coverage is *basic transmission loss* (L_b). L_b is the signal loss expected between ideal, loss-free, isotropic transmitting and receiving antennas [22]. This loss is a function of the frequency, pathlength, and attenuation on the path. The major source of attenuation in our survey area was obstruction of the radio path by buildings and vegetation.

Coverage is the percent of locations for which L_b does not exceed the allowable loss ($L_b^{\max}$) for a given system at the desired availability level. If one knows the operating parameters for a radio system, then an $L_b^{\max}$ can be determined based on the available transmitter power and the necessary SNR at the receiver to achieve the required bit error rate (BER). A station then has coverage if $L_b \le L_b^{\max}$.

Availability is based on the time variability of the received signal measured at each station. The cumulative

distribution function (CDF) of the received power is used to calculate the time statistics (i.e., availability) of L_b. For instance, the median signal power measured at a receiver station gives L_b for 50% availability, while the lower decile gives L_b for 90% availability. Using L_b calculated at specific availability levels and the statistical development of the previous section, area coverage and standard error estimates are made for a range of $L_b^{\max}$. As one would expect, coverage decreases at increased availability levels. Because a high level of availability is desirable, we have summarized area coverage results versus $L_b^{\max}$ assuming 99% availability. For coverage estimates at higher availability levels, more independent measurements would be required.

Because coverage estimates for both Northglenn and San Jose are similar, a sample of results from both sites are used to illustrate the general trends. Area coverage for a 0.5-km cell quadrant and a 1.0-km cell quadrant, both using a 12-m-high transmitter site in San Jose, are shown in Fig. 8. From the figure we can see that systems capable of sustaining an $L_b^{\max}$ of 150–155 dB can achieve 80% coverage at 99% availability in 0.5-km cell quadrants (1-km transmitter spacing). For the 1-km quadrant (2-km transmitter spacing), the coverage for L_b between 150 and 155 dB decreases to 75% versus 80% measured in the 0.5-km quadrant. In Fig. 9, we see that the area coverage for San Jose is improved significantly for the 0.5- and 1-km quadrants by using a 24-m-high transmitter site. Now, 80% coverage for the 0.5-km quadrant can be achieved at an L_b of 140 dB, 10–15 dB less signal loss than the 12-m transmitter results. When using a 24-m-high transmitter in the 1-km quadrant, 80% coverage can be reached at an L_b between 145 and 150 dB. For the 2-km quadrant we see that 80% coverage is not achieved for an L_b up to 155 dB. More details can be found in Ref. 23.

4.6. Attenuation

Attenuation is the additional power loss above the free space loss (spreading loss) between the transmit and receive antennas. It is convenient to separate L_b into its two components, attenuation A and basic free-space loss L_{fs}:

$$L_b(\text{dB}) = L_{fs}(\text{dB}) + A(\text{dB}).$$

Using this relationship, A is calculated by subtracting L_{fs} from L_b where L_{fs} is

$$L_{fs}(\text{dB}) = 32.4 + 20\log f(\text{MHz}) \cdot d(\text{km}).$$

An attenuation versus distance graph for San Jose using the 12-ft transmitter site is shown in Fig. 10. The data is highly scattered due to the random nature of the obstructions. However, a general trend can be seen by overlaying a linear least squares fit curve on the data. Similar linear fits were made using the other Northglenn and San Jose data. The slopes and intercepts of these curves are summarized in Table 5.

The slope of the attenuation data has an expected inverse correlation with the area coverage results. Northglenn, which had a smaller coverage estimate

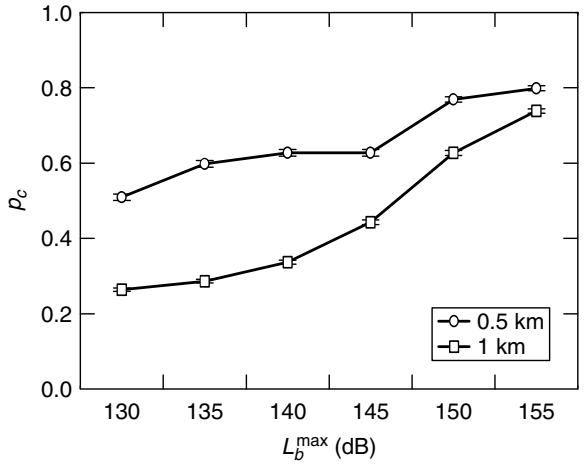


Figure 8. Area coverage estimate p_c versus $L_b^{\max}$ at 99% availability for 0.5 km and 1.0 km cells using a 40-ft transmitter, San Jose, California.

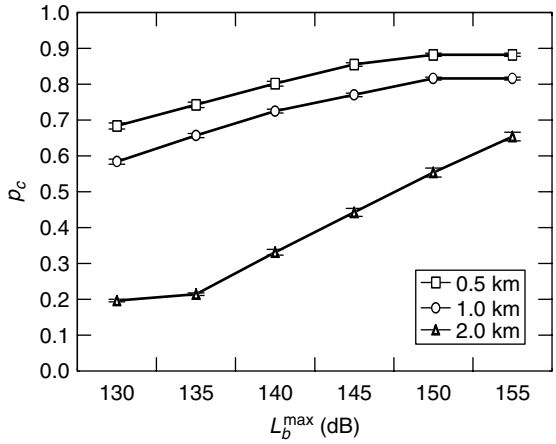


Figure 9. Area coverage probability estimate p_c versus $L_b^{\max}$ at 99% availability for 0.5-, 1.0-, and 2.0-km cells using the 80-ft transmitter, San Jose, California.

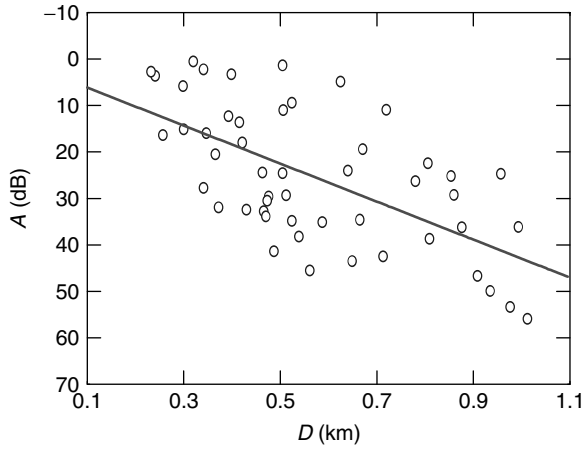


Figure 10. Attenuation versus distance using the 40-ft transmitter, San Jose, California.

Table 5. Attenuation versus Distance Slope and Intercept Data for Northglenn, CO and San Jose, CA

Site	Slope (dB/km)	Intercept (dB)
Northglenn (40-ft transmitter)	42.4	7.3
San Jose (40-ft transmitter)	40.9	2.1
San Jose (80-ft transmitter)	6.7	9.4

than San Jose, has the larger attenuation slope. The attenuation slope decreased significantly when the 24-ft transmitter site was used in San Jose, indicating that the radio path was able to clear many more obstructions. When a tree is blocking the radio path, signal propagation will be dependent on scattering and diffraction. In many cases when a tree was obstructing the radio path it was located within 10–20 m of the receiver site. For the 500-m cell, using an average pathlength of 250 m and assuming an obstruction (e.g., tree) at 235 m, the diameter of the first Fresnel zone for a 30 GHz signal is about 1 m. Usually LoS radio links require 60% of the first Fresnel zone to be free of obstructions to limit diffraction losses [22]. Hence, at least a 77-cm opening through the tree canopy is required for an unobstructed radio path.

In addition to arguments using Fresnel diffraction zones, large signal attenuation by trees is consistent with previous experiments to characterize millimeter wave propagation in vegetation [24–27]. Measurements in regularly planted orchards have found attenuation values between 12 and 20 dB per tree for one to three deciduous trees and up to 40 dB for one to three coniferous trees. The measured attenuation can be accounted for by a combination of one to four coniferous or deciduous trees on the radio path.

4.7. Cross-Polarization Discrimination

Cross-polarization discrimination measurements were made to test the practicality of frequency reuse schemes that employ signals of orthogonal polarization. A vertical linearly polarized signal was transmitted and both vertically and horizontally polarized signals were received. The larger XPD is, the more effective orthogonal frequency reuse will be. At millimeter-wave frequencies, rain-induced depolarization is produced by differential attenuation caused by nonspherical raindrops. As discussed previously, the effects of rain-induced depolarization for a short 30-GHz terrestrial link is expected to be small. Of more concern is the depolarization caused by scattering from vegetation. Experiments [24–27] have characterized millimeter wave depolarization in both coniferous and deciduous orchards. The most serious impairments are seen consistently in conifer tree stands where the average XPD at 28.8 GHz was 12 dB for foliage depths of 20 m and decreased to about 9 dB after 60 m. However, it is difficult to apply these results to cells proposed for LMDS applications because the foliage depth and tree species for any particular subscriber are random and unknown. Measured XPD results for Northglenn are presented as a function of attenuation in Fig. 11. The data are highly scattered but

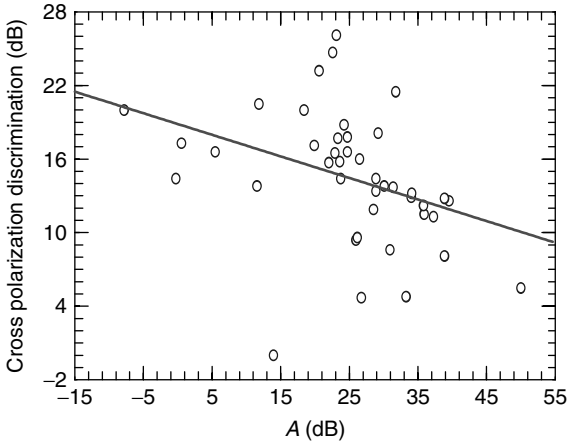


Figure 11. Cross-polarization discrimination (XPD) versus attenuation for 0.5-km cells using 40-ft transmitters in Northglenn, Colorado.

a linear fit predicts an XPD of 14 dB at an attenuation of 30 dB, which is 10 dB greater than predicted due to rain.

4.8. Characterization of Multipath Using the Tapped Delay Line Channel Model

The tapped delay line channel model is

$$h(t) = \sum_{n=1}^{N} \beta_n \delta(t - \tau_n) e^{-j\omega_c \tau_n}$$

where $h(t)$ is the complex channel impulse response, N is the maximum number of taps, n is the tap index, β is the tap gain, τ is the tap delay, and ω_c is the carrier frequency.

We selected three stations located at successively greater distances from the transmitters along the same cell radial to represent good, moderate, and bad wideband channels. Table 6 summarizes the channel model at these stations. The small delay spreads confirm that there are few specular reflections due to the filtering effect of the narrow beam receiver antennas. We note that delay spread is calculated using a 20-dB threshold. Table 7 lists the distance (D) between transmitter and receiver, attenuation (A), delay spread (S) and L_b for these paths. From Table 4 we see that links that exhibit multipath also have larger values of L_b and attenuation. Delay spreads are also plotted versus attenuation in Fig. 12. This plot also indicates that multipath is associated with larger signal attenuations.

5. SUMMARY

An LMDS band allocation was established by the FCC to provide broadband wireless access services to MANs and LANs. The spectrum allocation straddles the EHF and SHF bands near 30 GHz. Radiowaves in this part of the spectrum are commonly called millimeter waves. Although the radio bands were allocated and some technical specifications were made by the FCC, interference and coexistence rules were not defined for the broadcast

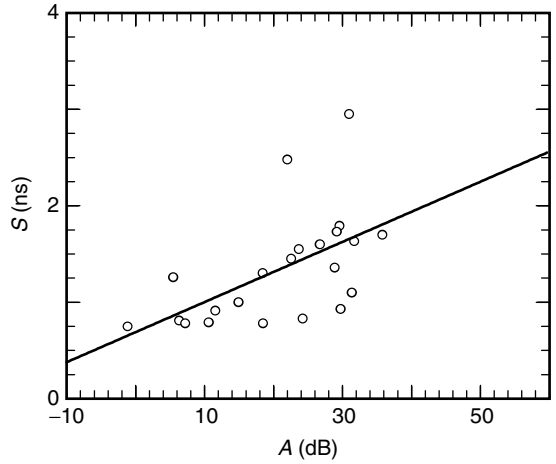


Figure 12. Delay spread (S) versus attenuation, 40-ft transmitters, Northglenn, Colorado.

Table 6. Summary of Tapped Delay Line Models for Good, Moderate, and Bad Channels from Northglenn, CO

Quality	Tap #	β_n (dBm)	τ_n (ns)
Good	1	0	0
Moderate	1	0	0
Moderate	2	−13.7	5.3
Bad	1	0	0
Bad	2	−2.8	3.6
Bad	3	−16.2	15.3

Table 7. Summary of Distance D, Attenuation A, Delay Spread S, and Basic Transmission Loss L_b at 99% Exceedance for Three Wideband Channels in Northglenn, CO

Quality	D (m)	A (dB)	S (ns)	L_b (dB)
Good	122	6.2	1.26	111.7
Moderate	309	32.2	1.60	145.9
Bad	419	32.6	2.95	159.4

mode of operation. To define these standards the IEEE LAN/MAN standards group 802.16 was established. This group is expected to publish physical layer (PHY) and media access control (MAC) Layer standards established by a consortium from private industry and government.

The advantages and disadvantages of LMDS are related to the use of the millimeter wave portion of the radio spectrum. The millimeter wave spectrum allows some equipment miniaturization and has large available bandwidths necessary for high-speed digital communication. But there are also significant problems associated with millimeter wave systems such as radiowave propagation impairments, the higher cost of electronic components, and unavailability of high-power solid-state linear amplifiers.

Radiowave propagation considerations for point-to-point links include attenuation caused by rain and atmospheric adsorption and depolarization due to nonspherical raindrops. These effects can be estimated using models and empirical formulas. Advanced computer models that incorporate high resolution areal photography and digital terrain data can be used to determine LoS paths excluding blockage due to vegetation. For broadcast systems, measurement data for specific environments must be used to determine coverage, availability levels, and radio channel characteristics such as multipath and signal depolarization. To date only limited measurement data are available at millimeter wave frequencies.

BIOGRAPHIES

Peter B. Papazian (M'91) received his B.S. in physics from the State University of New York at Stonybrook in 1973, and his M.S. in geophysics from the Colorado School of Mines in 1979. In 1990 he joined the radio research and standards group at the Institute for Telecommunication Sciences in Boulder, Colorado. At ITS, Peter has conducted research in the fields of millimeter-wave propagation, man-made radio noise, and impulse response measurements and systems. Currently, Mr. Papazian has developed an advanced antenna test-bed to study the capacity of data and mobile communication systems.

Roger A. Dalke received his bachelors degree in physics from the University of Colorado in 1971. He received his M.S. in geophysics in 1983 and his Ph.D. in 1986 from the Colorado School of Mines. As a research engineer, he has developed numerical techniques for a variety of electromagnetic scattering problems as well as signal processing and imaging methods used in exploration geophysics. More recently, he has been involved in the development of computer simulation models for digital radio systems, noise and interference measurements and analysis, and radio propagation in urban environments.

BIBLIOGRAPHY

1. R. H. Abrams, B. Levush, A. A. Mondelli, and R. K. Parker, Vacuum electronics for the 21st century, *IEEE Microwave Mag.* 61–72 (Sept. 2001).
2. FCC 96-311, *First Report and Order and Fourth Notice of Proposed Rule Making*, Docket 92-297, adopted July 17, 1996.
3. FCC 97-82, *Second Report and Order, Order on Reconsideration, and Fifth Notice of Proposed Rule Making*, Docket 92-297, adopted March 11, 1997.
4. *Code of Federal Regulations*, Title 47: Telecommunication, Section 101.103 to 101.113, Office of the Federal Register, National Archives and Records Administration, Oct. 1, 2000.
5. H. J. Liebe, MPM — an atmospheric millimeter-wave propagation model, *Int. J. Infrared Millimeter Waves* **10**: 631–650 (1989).
6. H. J. Liebe, G. A. Hufford, and M. G. Cotton, Propagation modeling of moist air and suspended water/ice particles at frequencies below 1000 GHz, *Proc. AGARD Conf. Atmospheric Propagation Effects through Natural and Man-Made Obscurants for Visible to MM-Wave Radiation*, 1993, pp. 3-1–3-11.

7. ITU-R (International Telecommunication Union, Radiocommunications Assembly), *Attenuation by Atmospheric Gases*, Rec. ITU-R P.676-4, Geneva, Switzerland, 1999.

8. R. H. Espeland, E. J. Violette, and K. C. Allen, *Atmospheric Channel Performance Measurements at 10 to 100 GHz*, NTIA Report 84-149, Apr. 1984 (NTIS Order PB 84-211325).

9. L. J. Ippolito, *Radiowave Propagation in Satellite Communications*, Van Nostrand Reinhold, New York, 1989.

10. R. L. Olsen, D. V. Rogers, and D. B. Hodge, The aR^b relation in the calculation of rain attenuation, *IEEE Trans. Antennas. Propag.* **AP-28**: 318–329 (March 1978).

11. K. C. Allen, *EHF Telecommunication System Engineering Model*, NTIA Report 86-192, April 1986 (NTIS Order PB 86-214814/AS).

12. E. J. Dutton, C. E. Lewis, and F. K. Steele, *Climatological Coefficients for Rain Attenuation at Millimeter Wavelengths*, NTIA Report 83-129, Aug. 1983 (NTIS Order PB 84-104272).

13. ITU-R (International Telecommunication Union, Radiocommunications Assembly), *Specific Attenuation Model for Rain for Use in Prediction Methods*, Rec. ITU-R P.838-1, Geneva, Switzerland, 1999.

14. R. K. Crane, Prediction of attenuation by rain, *IEEE Trans. Commun.* **COM-28**: 1717–1733 (Sept. 1980).

15. ITU-R (International Telecommunication Union, Radiocommunications Assembly), *Propagation Data and Prediction Methods Required for the Design of Terrestrial Line-of-Sight Systems*, Rec. ITU-R P.530-8, Geneva, Switzerland, 1999.

16. R. K. Crane, Comparative evaluation of several rain attenuation prediction models, *Radio Sci.* **20**(4): 843–863 (July–Aug. 1985).

17. E. J. Dutton and F. K. Steele, *Some Further Aspects of the Influence of Raindrop-Size Distributions on Millimeter-Wave Propagation*, NTIA Report 84-169, Dec. 1984 (NTIS Order PB 85-168334).

18. P. L. Rice and N. R. Holmberg, Cumulative time statistics of surface-point rainfall rates, *IEEE Trans. Commun.* **COM-21**: 1131–1136 (Oct. 1973).

19. E. J. Dutton and H. T. Dougherty, Year-to-year variability of rainfall for microwave applications in the U.S.A., *IEEE Trans. Commun.* **COM-27**(5): (May 1979).

20. National Oceanic and Atmospheric Administration, *Climates of the States*, 1978, Vol. 1, p. 136.

21. M. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Macmillan, New York, 1977.

22. M. P. M. Hall, Effects of the troposphere on radiocommunication, in *IEE Electromagnetic Wave Series 8*, Peter Peregrinus Ltd., Stevenage, UK and New York, 1979, pp. 10–13, 81.

23. P. B. Papazian and G. A. Hufford, Study of the local multipoint distribution service radio channel, *IEEE Trans. Broadcast.* **43**(2): (June 1997).

24. D. Jones, R. Espeland, and E. Violette, *Vegetation Loss Measurements at 9.6, 28.8, 57.6, and 96.1 GHz through a Conifer Orchard in Washington State*, NTIA Report 89-251, Oct. 1989 (NTIS Order PB 90-168717).

25. E. Violette, R. Espeland, and F. Schwering, Vegetation loss measurements at 9.6, 28.8, and 57.6 GHz through a pecan orchard in Texas, in *Multiple Scattering of Waves in Random Media and Random Rough Surfaces*, Pennsylvania State Univ., State College, PA, 1985, pp. 457–472.

26. P. B. Papazian, D. Jones, and R. Espeland, *Millimeter-Wave Propagation at 30.3 GHz through a Pecan Orchard in Texas*, NTIA Report 92-287, Sept. 1992.

27. E. Violette, R. Espeland, and K. C. Allen, *Millimeter-Wave Propagation Characteristics and Channel Performance for Urban-Suburban Environments*, NTIA Report 88-239, Dec. 1988 (NTIS Order No. PB 89-180251/AS).

28. R. A. Dalke, G. A. Hufford, and R. L. Ketchum, *Radio Propagation Considerations for Local Multipoint Distribution Systems*, NTIA Report 96-331, Aug. 1996.

LOCAL AREA NETWORKS

JOHN H. CARSON
George Washington University
Washington, District of Columbia

1. INTRODUCTION

Local area networks (LANs) are private, high-speed networks that are limited in distance and typically serve as a distribution system for both Internet and local information services.

2. HISTORY

Although a number of research and development activities can be associated with the origin of the LAN, the best known early publication in this field appeared in 1976 [1] by Robert Metcalfe and David Boggs, who developed the Ethernet Local Area Network system while working at Xerox PARC. This coaxial cable-based system transmitted data at 2.94 Mbps. Following development of the Ethernet LAN, other local network technologies appeared and disappeared, most notable of which was the token ring system developed by IBM.

In 1980, Xerox, Digital Equipment Corporation, and Intel developed the "Ethernet Blue Book" or "DIX standard." The second version of this standard was completed in November 1982. Also in 1980, the IEEE formed the 802 Committee to standardize LAN/MAN (metropolitan area network) technology. This committee continues to develop and extend LAN standards.

2.1. Xerox PARC

As mentioned above, Robert Metcalfe and David Boggs published the details of Ethernet, a project developed at the Xerox Palo Alto Research Center in 1976 (Fig. 1). Although developed in 1976 the concepts presented in this paper are still the foundation for the contention based LANs of today.

It should be noted that Metcalfe later founded 3COM Corporation, which was instrumental in transitioning Ethernet from the laboratory to the commercial marketplace. Thus he developed the concept in a research environment and then guided its transition to the commercial world.

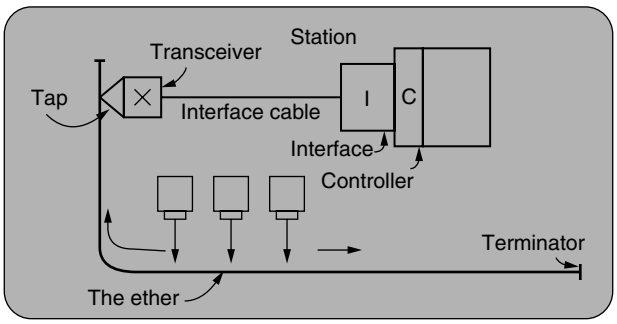


Figure 1. Diagram presented at the 1976 National Computer Conference by Robert M. Metcalfe.

2.2. 802 Activities

In 1980, noting the increasing popularity of Ethernet and the need to standardize existing and future LAN protocols, the IEEE formed the 802 Committee, whose duties were to oversee these standards. Later, the Committee's responsibilities were increased to address MAN standards. As can be seen by the sample of IEEE 802 working group activities listed below, this effort has been quite comprehensive (see Table 1).

As the standards were developed, many transitioned to ANSI and eventually ISO standards. Notable among these standards are 802.1, 802.2, 802.3, 802.4, 802.5, and 802.11. IEEE 802.1 defines the overall architecture of the set of standards; 802.2 defines the logical link control (LLC), which provides a standard set of network services to higher-level protocols; and 802.3, a contention-based system, addresses the copper-based Ethernet standards, which are still employed. 802.3 has expanded from the original 10-Mbps coaxial cable standard to one employing twisted pairs and fiber optics operating at 10, 100, 1000, and 10,000 Mbps.

While the 802.3/Ethernet standard describes a contention-based system, the 802 Committee has developed several contention-free, token passing systems, the most significant of which is the 802.5 — a token ring system popularized by IBM that has practically become extinct. 802.4, a token bus system designed primarily for/by the automotive industry, never achieved popularity and is also extinct.

Table 1. Sample of IEEE 802 Activities

P802.1, *High Level Interface* (HILI)	P802.2, *Logical Link Control*[a]
P802.3, *CSMA/CD*	P802.4, *Token Bus*[a]
P802.5, *Token Ring*[a]	P802.6, *Metropolitan Area Network* (MAN)[a]
P802.7, *Broadband TAG*[a]	P802.8, *Fiber Optic TAG*[b]
P802.9, *Integrated Services LAN* (ISLAN)[a]	P802.10, *Standard for Interoperable LAN Security* (SILS)[a]
P802.11, *Wireless Local Area Network* (WLAN)	P802.12, *Demand Priority*[a]
P802.14, *Cable-TV Based Broadband Communication Network*[a]	P802.15, *Wireless Personal Area Network* (WPAN)
P802.16, *Broadband Wireless Access*	P802.17, *Resilient Packet Ring*

[a]Inactive.
[b]Disbanded.

The token passing systems avoid contention by having the stations organized in a ring (physical for 802.5 and logical for 802.4). A token is passed from station to station. If it has nothing to send, the station receiving the token will directly pass the token to the next station or if it does have packets queued for delivery, it will send one or more of those packets before passing the token to the next station. This deterministic behavior offers capabilities not available in the contention-based 802.3 systems, such as supporting priority schemes and providing worst-case response times. However, these features were not adequate to overcome the overwhelming popularity of the Ethernet systems, and they became extinct as hub-based 802.3 systems eliminated the capability gap between the contention and token-based systems.

As the 802.3 hubs increased in sophistication, they began to use alternative technologies internally. This move changed the role of the IEEE standards from describing the overall activity and behavior of the network to that of describing the *network interface* to the hub. 802.11 covers a set of wireless Ethernet standards, including 802.11b, which is currently commercially popular and discussed in detail later in this article.

3. ETHERNET FUNDAMENTALS

Contention based LAN systems employ a broadcast approach where every station potentially hears every transmission. This means that overlapping transmissions (from different stations) will *collide* and interfere with each other. In order to avoid collisions or quickly recover from those collisions not avoided, the CSMA/CD concept is employed. Originally known as *listen before talk*, and now known as *carrier sense multiple access* (CSMA), the first part of this approach requires each station to listen to the medium and detect the presence of any transmission. If no transmission is detected, then the station may go ahead and transmit. If the medium is busy, the station defers until the medium becomes free. In order to allow transmission detection, the modulation scheme employs a carrier that is quickly distinguishable from a quiescent state. For the original Ethernet and 802.3 standards differential Manchester encoding was employed. This modulation scheme requires a minimum of one line state transition per bit, making it easy to distinguish from a quiescent line.

However, CSMA, by itself, is not sufficient to avoid collisions. Multiple stations could simultaneously sense an empty medium and decide to transmit at roughly the same time, thereby creating a collision. In order to operate efficiently, LANs must detect and quickly recover from collisions since (1) the time involved in a collision is wasteful; and (2) since the messages are obliterated, they become lost frames and thus require some action at higher levels (TCP in the Internet), which noticeably

degrades performance. Originally known as *listen while talk*, *collision detection* (CD) is handled a number of ways. In coaxial cable–based networking, the transmitting station listens to the network while transmitting. If the message observed is not identical to that being transmitted, a collision has occurred. At this point, the detecting station continues to transmit for a short jam-time period (or alternatively sends a *jamming* signal for that same interval) in order to allow the collision to be noticed by all involved parties.

Twisted-pair technologies use hubs that employ separate pairs for transmitting and receiving data. When receiving a transmission, the hubs relay it to all connected stations *except* the originating station. Thus, if a transmitting station hears an incoming transmission, a collision is taking place because the transmission has originated from a different station.

When a collision is detected by whatever means employed, the participating stations transition into a backoff state. Essential to this technique is the concept of a *slot time*. The value of a slot time varies with the implementation standards, but it must satisfy the following criteria [2]:

- *It must define an upper bound on the acquisition time for the medium.*
- *It must define an upper bound on the length of a frame fragment generated by a collision.*
- *It is used for scheduling retransmissions.*

The first two criteria dictate that the slot time will be at least equal to the longest round-trip propagation delay between two stations on the same LAN plus the jam time. This propagation delay involves the signal propagation through the medium plus any electronic delays induced by hubs, repeaters, and level 2 switching. More specifically, it is the longest time period for which a transmitting station must transmit before being assured that a collision will not take place.

The backoff technique employed by IEEE 802.3 and Ethernet is *truncated binary exponential backoff*. Here, when encountering a collision, each station waits until the medium is clear (the collision has ended); it then waits an integer number of slot times chosen from a randomly distributed set of integers in a specified range. If a collision again occurs, the integer range is increased. Eventually, the station either transmits successfully or gives up after a backoff attempt limit and declares a system failure.

The specific integer range for the nth transmission retry is $0 \leq r \leq 2^k$, where $k = \min(n, 10)$.

The set of stations that may enter into a collision or see a collision fragment is known as a *collision domain* while the set of stations that can receive a single broadcast is known as a *broadcast domain*. These are often the same two sets, but the collision domain may be a proper subset of the broadcast domain through the use of switching hubs and other devices that pass broadcast messages but block collisions.

4. IEEE STANDARDS

4.1. IEEE Model

As mentioned earlier, the IEEE 802 Committee has developed both an architecture and an associated set of protocols that are quite thorough. Unfortunately, a complete explanation would take longer than space here allows. Therefore, only an overview is provided here. The IEEE architecture is shown in Fig. 2.

The IEEE model addresses two major ISO model layers: the physical and the data link layers. The ISO physical layer corresponds closely to the IEEE physical layer, while the ISO data link layer contains two sublayers: the IEEE media access control (MAC) and logical link control (LLC) layers.

Multiple link service access points (LSAPs) provide LAN services to higher level layers. Unacknowledged connectionless (type 1), connection-oriented (type 2), and acknowledged connectionless (type 3) are defined in the standard. These LSAPs also hide the MAC/PHY level differences between the various options. IEEE 802.3 (the Ethernet style) employs the type 1 (unacknowledged connectionless) service.

The MAC sublayer supports the LLC sublayer by providing the necessary functions for the LLC to perform. Specifically, it provides for the transmission and reception of frames, which involves framing, addressing and frame check sequence generation and checking.

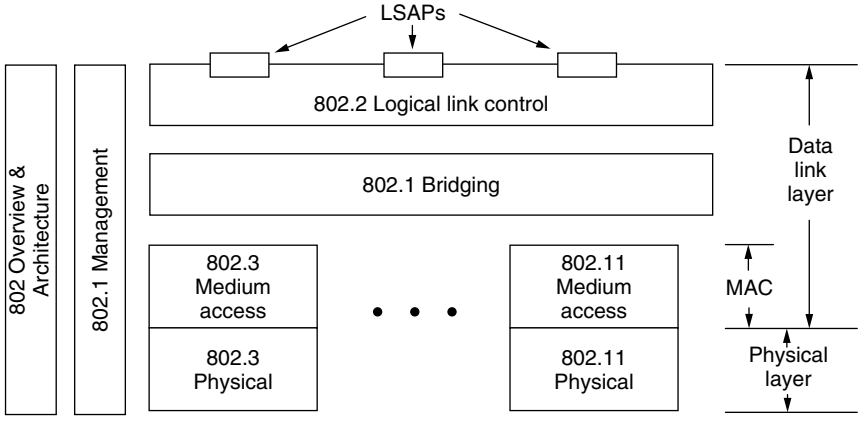


Figure 2. IEEE 802.11 standards.

The physical layer deals with the actual transmission and reception of signals, modulation and timing issues, and so on.

4.2. IEEE Address Management

The developers of Ethernet recognized the value of ensuring that each LAN device, regardless of the PHY layer, has a unique and standard address. In reality, this is only critical for devices on the same LAN since traffic across different LANs is handled by Internet protocols. Originally, Xerox administered LAN addresses, but now IEEE holds this responsibility.

Known as *MAC addresses*, LAN addresses are 48 bits in length (6 octets). The first 2 bits of the address define its nature. The first bit indicates whether the address is a unique (set to 0) or multicast address (set to 1). The second bit, if set to 0, indicates the address is administered by the IEEE. If set to 1, the address is locally administered and not subject to any of the following specifications. Obviously, virtually all LAN addresses are globally administered by the IEEE.

The IEEE addresses are further split into 24-bit portions. The first 24 bits define the *organizationally unique identifier* (OUI), which is administered by the IEEE and allocated uniquely to requesting organizations. Thus it is impossible for LAN addresses specified by one vendor (employing that vendor's OUI) to duplicate addresses from another vendor. Each vendor is responsible for avoiding duplication within its OUI address space. Each distinct OUI allows the vendor to develop approximately 4 million group and unique addresses. If an organization does exhaust its address space, it simply applies for another OUI. Duplication of IEEE addresses has been observed but attributed to manufacturing defects.

4.3. IEEE Media Access Control Frame Format

Originally, the IEEE 802.3 and Ethernet standards defined slightly different MPDU (MAC) frame formats as shown in Fig. 3.

The original differences between the IEEE 802.3 and Ethernet standards were minimal. The most significant was that the Ethernet standard employed a protocol type field to identify the protocol that either requested transmission of the frame or should receive the frame. The 802.3 standard employs the 802.2 layer between the 802.3 and IP layers. The 802.2 SNAP format presents an alternative means for providing the same identification. The 2-octet length/type field is used to distinguish between the approaches. Values less than or equal to 1500 indicate the frame is an IEEE 802.3 frame and the field is a length field containing the length of the remainder of the frame while values above or equal to 1536 represents a protocol id (e.g., 2048 indicates Internet IPv4) and an Ethernet frame. In 1997, revisions to 802.3 merged the Ethernet format into 802.3 as an option.

As shown in Fig. 4 the updated frame format merges the two fields into a single length/type field.

When virtual LANs (VLANs) appeared, the Ethernet frame format was again modified as discussed later. The preamble is a sequence of 56 bits that begins with 1, alternating zeros and ones, and ends with 0. This pattern allows the receiver to synchronize with the incoming data. The start of the important frame contents is indicated by a *start frame delimiter* containing 10101011. Following the start frame delimiter are the six-octet destination and source MAC addresses, and following the MAC addresses is the length/type field previously mentioned. The pad field is used to ensure that the MAC frame size meets the minimum requirements of the 802.3 MPDU. This minimum size (the number of octets beginning with the source address and including everything through the 32 octet FCS) varies with the particular 802.3 implementation — for example, 64 octets for the 10BASE options.

4.4. IEEE PHY Level

The Physical (PHY) Level provides the capability of transmitting and receiving bits between Physical Layer Entities [3] through the defined modulation and encoding schemes specified in the standard.

802.3:

Preamble (7 octets)	Starting delimiter (1 octet)	Destination address (6 octets)	Source address (6 octets)	Length (2 octets)	802.2 Frame (0 – n octets)	Pad	FCS (4 octets)

Ethernet:

Preamble (8 octets)	Destination address (6 octets)	Source address (6 octets)	Type field (2 octets)	Data (46 – 1600 octets)	FCS (4 octets)

Figure 3. Original 802.3/Ethernet MPDU organization.

Preamble (7 octets)	Starting delimiter (1 octet)	Destination address (6 octets)	Source address (6 octets)	Length/ Type field (2 octets)	MAC client data	Pad	FCS (4 octets)

Figure 4. Final 802.3 MPDU frame format.

4.5. IEEE 802.3: Current Copper LAN Implementations

Through the years, the IEEE has developed a large number of LAN standards, originally designated as follows:

<data rate in Mb/s> <medium type> <maximum segment length (× 100 m)>

For example, 10BASE5 would signify a 10-Mbps baseband system with a maximum cable length of 500 meters.

Later options dropped the maximum segment length portion for a medium designation such as "T." Thus, 10BASE-T would signify a 10-Mbps baseband twisted-pair system. The standards began with coaxial cable-based systems and then moved to twisted-pair systems also increasing the data rate. Table 2 lists the most common options.

4.5.1. 10BASE5. 10BASE5, the oldest member of the IEEE 802.3 effort, is now mostly extinct. A 10-Mbps descendant of Metcalfe's Xerox PARC system, 10BASE5 used a thick 50-Ω coaxial cable (polyvinyl chloride with a 0.40-in. diameter or fluoropolymer with a 0.37-in. diameter). 10BASE5, which allowed cable spans of up to 500 m, is the only IEEE 802.3 standard that employed an external medium attachment unit (MAU) known as a *transceiver*. These transceivers connected to the Ethernet coax (coaxial cable) via vampire taps, which attach to the inner conductor through a hole in the outer layers of the coax. An AUI (transceiver) cable connected the transceiver to the station. A failure in the cable would take the entire coax-based system down, thereby making coaxial cable problematic. If the cable were bent, crushed or a terminator removed from either end, the signal quality would be degraded enough, due to reflections, for the system to fail.

4.5.2. 10BASE2. The 10BASE2 standard, also known as *Thin Ethernet* or *Cheapernet*, employs a 50-Ω, RG-58 A/U cable, which is smaller in diameter than the 10base5 cable. With this standard cable sections threaded their way through each of the stations by employing a T-tap at each station that brought the signal into the station. The maximum cable length for this standard was only 185 m as compared to 500 m for the 10BASE5. This did not pose a significant problem, however, since computers had become cheaper and more plentiful and the cable could reach enough computers to make it practical. One weakness was that each cable segment required two BNC connectors which were a common point of failure.

4.5.3. 10BASE-T. Although not the first twisted-pair LAN since 1BASE-T and other non-IEEE systems existed previously, the 10BASE-T quickly became the standard LAN technology employed. This was true in part because the installation and maintenance of twisted-pair cable plants was far easier than that of the coaxial cable systems. Additionally, RJ-45 connectors, which connect up to four twisted pairs, are used to connect the cables to hubs. The use of hubs allowed fault tolerance and, when managed with SNMP or a similar protocol, provided useful management and administrative capabilities.

Table 2. Popular 802.3 Options

10BASE5	Thick coaxial cable	10 Mbps
10BASE2	Thin coaxial cable	10 Mbps
10BASE-T	Two twisted pairs	10 Mbps
100BASE-TX	Two twisted pairs	100 Mbps
1000BASE-T	Four twisted pairs	1 Gbps

The 10BASE-T systems center around a hub (see Fig. 5), known originally in the IEEE standard as a multiport repeater. Unlike the two-way transmission possible with coaxial cable, the twisted-pair systems employ separate pairs for transmitting and receiving. Originally, the system was designed to operate successfully with Category 3 voice-grade unshielded twisted pair (Cat 3 UTP) cable. However, the improved performance of Cat 5 cable coupled with the emergence of 100BASE-T systems has eliminated Cat 3 cable from consideration for today's installations. The hub operates by reflecting any signals received on the uplink from a station to the downlinks for all but the transmitting station. Hubs may be spliced together by using either a special gender switching cable or a special repeater port on the hub. (Otherwise, the downlink from one hub would connect to the downlink of the other hub.)

The two-pair wire plant required a few changes to the CSMA/CD implementation. While transmitting, the 10BASE-T listens to the downlink rather than to the signal on the cable. If a signal appears on the downlink, it must have come from a station other than the listening station thus indicating a collision and thereby negating the need to compare the received message with the transmitted message. When a collision is detected, the detecting station continues to transmit and sends a *jamming* signal for the jam interval in order to inform all involved parties of the collision.

When the 10BASE-T system was introduced, vendors often had a mixed systems employing it and the two coaxial cable technologies. PC NICs (network interface cards) that provided all three interfaces (10BASE5, 10BASE2, and 10BASE-T) became common. Virtually all LAN standards that have appeared after 10BASE-T have employed either twisted pair cable or fiber optic cabling.

4.5.4. 100BASE-T. As technology advanced, network speed became increasingly important. The IEEE 802.3u

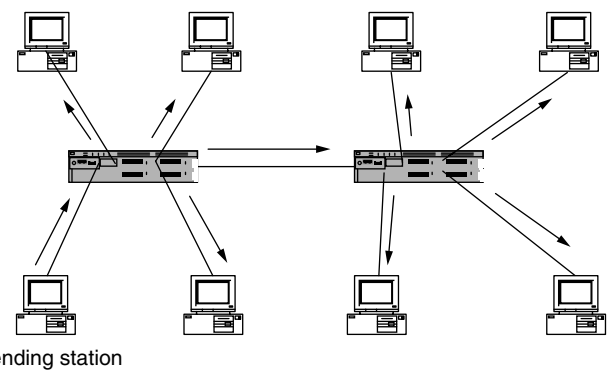


Figure 5. 10BASE-T hub system showing transmission path.

Table 3. 100BASE Options

TX — Two Cat 5 pairs
FX — Two optical fibers
T4 — Four Cat 3 pairs
T2 — Two Cat-4 pairs

Task Group was chartered with developing a 100 Mbps twisted-pair standard; a number of options were developed by the committee, with the 100BASE-TX the dominant option (see Table 3).

100BASE-TX does not differ significantly from 10BASE-T in that both systems can use Cat 5 cable plants. The physical modulation scheme does differ from 10BASE-T and follows that employed by *fiber distributed data interface* (FDDI). (The higher data rate makes differential Manchester encoding less practical as two transitions are often required for each bit.) Many components can automatically accommodate either 10BASE-T or 100BASE-TX through *autonegotiation*. This provides an effective upgrade path from 10 to 100 Mbps using Cat 5 cable plants.

4.5.5. Full-Duplex Ethernet. In 1997, IEEE 802.3x issued a standard for a full-duplex version of Ethernet that allows two stations to communicate over *point-to-point* links. It does not support hubs of other connections. Full-duplex transmission allows a maximum bandwidth of twice the conventional LAN since both transmission and reception can occur simultaneously. Additionally, since only two stations are involved, collisions do not occur, and therefore point-to-point links between the two stations are longer than that allowed in a true contention situation. PAUSE frames are added to the provide flow control necessary to support the higher bandwidth available in a contention free environment. The PAUSE frame specifies the amount of time that a receiving station must refrain from transmitting anything other than MAC control frames. Full-duplex Ethernet also allows *link aggregation* where multiple links (running at the same data rate) may exist between two stations. The bandwidths of the links may be combined to provide high bandwidth capability — for instance, two 100-Mbps full-duplex connections may be combined to provide a 200-Mbps connection.

4.5.6. 1000BASE. The IEEE Gigabit Ethernet standard can be logically divided into two groups. The first developed by the 802.3z taskforce, known as 1000BASE-X, contains three PHY options: 1000BASE-SX (short wavelength fiber) 1000BASE-LX (long wavelength fiber) and 1000BASE-CX (short-run copper). The second group, developed by the 802.3ab taskforce, 1000BASE-T, is designed as an extension to 10BASE-T and 100BASE-T.

The 1000BASE-X family uses the physical layer standards based on those employed by Fibre Channel technology. On the other hand, the 1000BASE-T standard uses four pairs of Cat 5 cable. (Note that 100BASE-T and 10BASE-T only use two pairs.) Each of the four pairs is modulated at the same clock rate (125 MHz) as 100BASE-T but employs a coding scheme that contains 2-bits/symbol and thus achieves 250 Mbps per twisted pair. Thus, four pairs transfer 1000 Mbps. As with the 100BASE-T standard, the maximum cable segment length is specified as 100 meters.

In order to meet the slot-time requirements of gigabit transmission, an extension field is added on to the end (after the FCS) of the Ethernet PDU. This is only employed in half-duplex operation as collisions do not occur in full-duplex mode. The extension bits are *non-data* symbols, which distinguish them from data bits.

For data rates above 100 Mbps, transmitting stations may employ a *burst mode* where a series of frames may be transmitted without relinquishing control of the transmission medium. *Burst mode* permits higher efficiency than would be possible with the conventional Ethernet protocol. The first frame may, if necessary, employ extension bits, but subsequent frames in the burst need not. Each frame is sent with the proper interframe gap, but the gap is filled with non-data symbols, which prevent the medium from appearing idle.

5. VIRTUAL LANs

Virtual LANS (VLANs) allow a set of stations to share the same broadcast domain while not necessarily in the same physical domain. A set of stations connected to a set of switches can be partitioned into several VLANs.

VLANs can be implemented by employing port-based VLAN hubs that partition stations based on the connection port. Connection ports are configured into VLANs through some form of station management, and then all stations plugged into the designed ports are in the same broadcast domain. Another approach is to partition based on MAC addresses. Both this and the previous approach do not require any special configuration of the stations.

VLANs implemented by using IEEE 802.1Q employ a tagged frame that contains a 4-octet tag inserted just after the source MAC address. The tag begins with 10000001 and contains a priority as well as a 12-bit VLAN identifier (the VID). This tagged frame is defined in the IEEE 302.3, 2000 edition standard (see Fig. 6).

6. WIRELESS LANs: 802.11

Wireless LANs, standardized by the IEEE 802.11 Working Group [4,5], extend the Ethernet concept to an RF connection to the desktop. As shown earlier in Fig. 2, the 802.11 medium-access and physical standards are

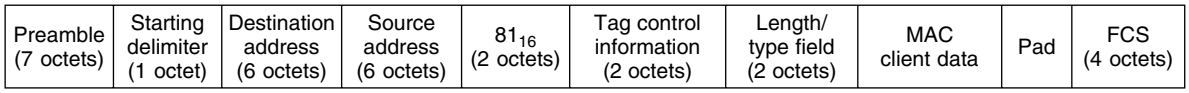


Figure 6. VLAN tagged frame format.

alternatives to the 802.3 or other 802 medium access and physical layer standards within the overall 802 architecture.

Designed to support mobile computing and improve flexibility of the "to the desktop" connection, wireless Ethernet has proved to be a very difficult technological chore. Issues such as multipath distortion, licensing, security, bandwidth, and possibly health hazards hindered the development of wireless systems. In 1997, the IEEE 802.11 Working Group published standards for both frequency-hopping spread spectrum (FHSS) and direct sequence spread spectrum (DSSS) 2.4 GHz and infrared technologies supporting data rates of 1 and 2 Mbps. Although some products appeared, they did not become commonplace until the 802.11b Task Group developed a 2.4 GHz direct sequence spread spectrum 11 Mbps system. The IEEE 802.11a Task Group also developed a 5 GHz orthogonal frequency division multiplexing system with data rates up to 54 Mbps. This technology is currently under development with the fabrication of integrated circuits that operate at this higher data rate.

In order to provide compatibility with existing wired LANs, the 802.11 Working Group designed 802.11 to provide the same interface as 802.3. Thus, it uses the 802.2 LLC sublayer and appears identical to 802.3 from any layer above 802.2.

Other wireless technologies that can be used for LANs, such as BlueTooth and HomeRF, are briefly addressed later in this section and covered in detail in separate entries in the encyclopedia.

6.1. 802.11 PHY

The 802.11b systems operate in the 2.4 GHz ISM (industrial, scientific, and medical) band, which does not require licensing. This band is split into 14 overlapping 22 MHz channels, but not all channels are available in all countries; for example, in the United States, only channels 1–11 are available. The power employed also varies by area, but most devices available in the United States limit the output power to 100 mW even though 1000 mW is the formal limitation. This puts the power output to less than that employed by the typical digital cellular phone (125 mW). Most systems also employ a speed backoff dropping from 11 Mbps to 5.5, 2, and then, finally, 1 Mbps.

The actual distances achieved through 802.11b are very sensitive to antenna placement, multipath distortion, walls, floors, and other barriers. Typical distances between components vary between 50 and 200 ft.

6.2. Network Topology

Due to the increased configuration flexibility and complexities associated with RF transmission, the overall architecture and protocols forming the 802.11 standard are more complex than their 802.3 counterparts. There are six major components of wireless LANs: the wireless stations themselves (STAs in IEEE terminology); access points (APs); the wireless medium; basic service sets (BSS); the distribution system (DS); and the extended service set (ESS) (see Fig. 7).

A set of stations (STAs), commonly called wireless NICs (network interface card), that talk to each other is called a *basic service set* (BSS). If all the STAs are wireless and there are no other components, the BSS is called an ad hoc or *independent BSS* (IBSS).

Access points (APs) serve both as relays between STAs and as bridges between STAs and wired LANs. When a BSS contains a single AP, it is called an *infrastructure BSS*, but the term "infrastructure" is typically omitted. In this mode, wireless stations communicate with each other with the AP acting as a relay. This slows STA-to-STA performance since two messages are needed for each STA-to-STA frame rather than the single STA-to-STA message employed in the ad hoc mode (see Fig. 8).

When a system contains more than one AP and associated BSS, it is called an *extended service set* (ESS). The BSSs are connected through an abstraction known as a *distribution system* (DS), which is typically a conventional copper-based LAN. The additional access points provide for increased geographic coverage in a manner similar to multiple cells for a cellular phone system.

6.3. 802.11 MAC Operation

The 802.11 MAC provides peer-to-peer best-effort, connectionless communication between LLC entities. Three types of MAC frames exist: data, control, and management (with a fourth type reserved for future use). MAC service data units (MSDUs) convey the peer-to-peer data supplied by high-level protocols, while MAC management PDUs

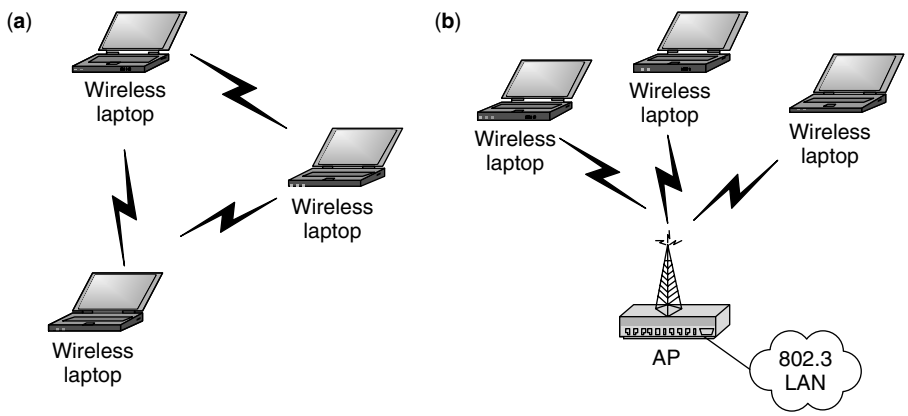


Figure 7. 802.11 architectures: (**a**) IBSS; (**b**) Infrastructure BSS.

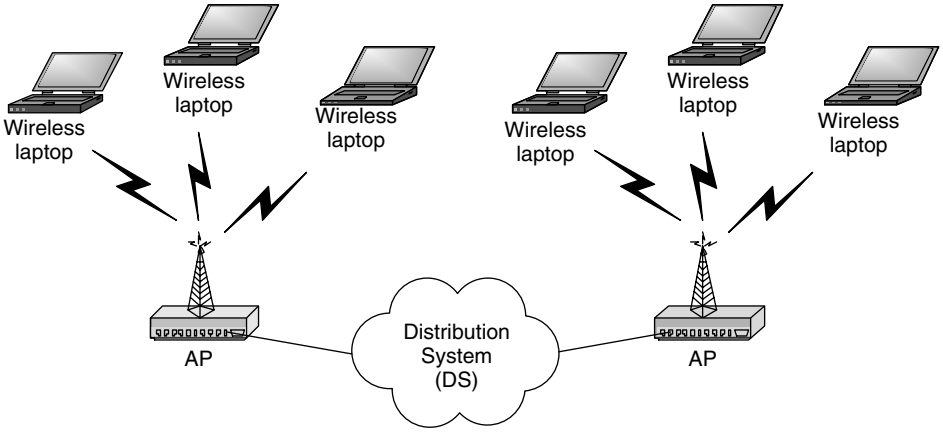


Figure 8. 802.11 Extended service set.

(MMPDUs) convey management messages supporting the communication. Additionally, control frames are employed to aid management and data transfer, and as with 802.3, multicast and broadcast services are available.

6.3.1. 802.11 MAC Data Frame Format. As shown in Fig. 9, the MAC data frame (MPDU) format differs from the 802.3 MAC data frame. MSDUs contain data supplied by a higher-level protocol and are one of the MPDU payload types. MMPDUs contain management information and are another MPDU payload.

MSDUs are prepended with a MAC header containing four 6-octet addresses, a frame control field, a duration/ID field, and a sequence control field. A 4-octet FCS (CCITT CRC-32) is appended to the MSDU.

The control field contains the 802.11 protocol version and a number of subfields to support fragmentation, power management, retries, WEP (*wired equivalent privacy* — discussed later), utilization, and other operational parameters.

Four 6-octet IEEE addresses are contained in the MPDU, although only some message types will use all four addresses; others will use between 1 and 3 addresses. In addition to the source and destination addresses found in 802.3, transmitter and receiver addresses as well as a BSSID address may be contained in the other two address fields. The transmitter address (TA) identifies the wireless source of the transmitted message (not always the originating source), while the receiver address (RA) specifies the wireless receiver (not always the ultimate destination.)

The duration/ID field contains information to update system NAVs or an association identificer (AID) used to obtained frames buffered in APs during STA power saving activities.

6.3.2. CSMA/CA. While conventional (wired) LANs employ CSMA/CD, wireless LANs cannot implement the collision detection (listen while talk), so instead they employ CSMA/CA (a collision avoidance scheme). Each STA employs the listen before talk aspect of the CSMA from 802.3. However, when the STA detects a busy medium, it defers for a time period determined by a binary exponential backoff algorithm. The range of values generated from this algorithm doubles each time the station consecutively defers due to a busy channel.

6.3.2.1. Hidden-Node Problem. A problem specific to wireless LANs is the *hidden node problem* (Fig. 10). Because of distance limitations of the RF system, two STAs may possibly communicate effectively with an AP

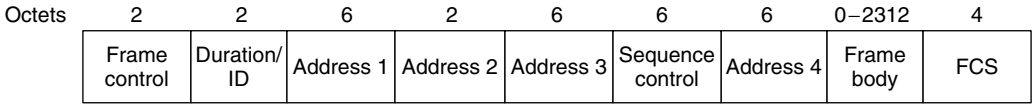


Figure 9. MPDU (MAC frame) format.

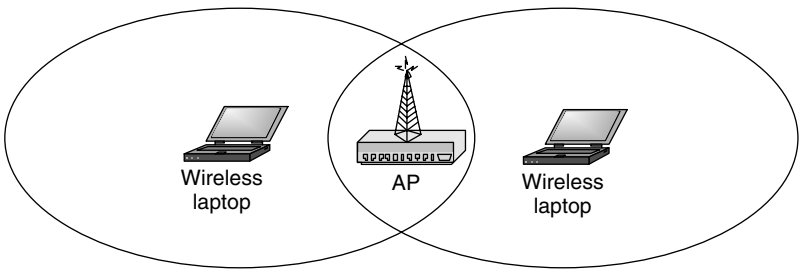


Figure 10. Hidden node problem.

but cannot hear each other. Thus, the CSMA algorithm cannot prevent collisions at the AP in all cases.

To work around this problem, STAs may send an *optional* RTS (request to send) message to the AP. Whether this RTS message is sent or not depends on the value of the dot11RTSThreshold attribute. If the frame is longer than this value, then a RTS message contains an appropriate reservation time period, which the access point echoes back to all stations in a CTS (clear to send) message. Then, the requesting station is clear, for the reserved time period, to send its frame without the possibility of interference from a hidden STA. A zero setting for dot11RTSThreshold requires all frames to be sent with RTS/CTS exchange, while a value of dot11RTSThreshold greater than the largest MSDU deactivates the RTS/CTS facility.

6.3.2.2. The Network Allocation Vector (NAV). Additional help for the collision avoidance system is provided through the NAV (network allocation vector), which indicates to a station the time period before the network medium becomes free again. NAVs are updated from data contained in each transmitted frame.

6.3.3. Roaming. When a STA enters the area of one or more APs, it chooses an AP (joining a BSS) based on signal strength and error rates. Joining is called *association*. When a STA shuts down, it *disassociates* with the system. *Reassociation* occurs when a station requests to switch associations between APs. Similar to the association service, this request also includes the AP previously involved in an association. The protocols to support reassociation coordination between APs are not standardized at this time, so APs from different vendors may not perform handoffs successfully.

6.3.4. Fragmentation. In order to increase system reliability, IEEE 802.11 allows transmitters to fragment unicast MSDUs and MMPDUs into smaller MPDUs. Defragmentation occurs when the frame arrives at the immediate receiving station All the fragments, each of which is acknowledged separately, are the same size except for the last, which may be smaller.

The value in the Sequence Number field is the same for all fragments of a single MSDU or MMPDU. The sequence order within the set of fragments is designated by the value in the Sequence Control field.

6.3.5. Power Management. To support power conservation in battery-powered equipment, STAs may employ a power management scheme. For an infrastructure BSS, a STA will be in one of two modes of operation. When a station is *awake*, it is fully powered and operational. When a STA is in the *doze* state, it is not able to transmit or receive. STAs notify their associated APs when they are changing state. If a STA enters the *doze* state, the AP must buffer any transmissions to it until the STA changes to the *awake* mode.

For an independent BSS, the operation is more complicated since there are no APs to buffer frames for dozing STAs. A dozing STA must periodically awake to allow any STA with a queued message for it to send the message. After draining the queued messages, the STA may again doze for a time period.

6.4. Security

The 802.11 standard identifies two security subsystems: the *wired equivalent privacy* (WEP) system and the *authentication services*. The goal of WEP is to provide security equivalent to that of a wire-based system. Thus, rather than being a true security system, WEP simply alleviates weaknesses introduced by the wireless nature of the network.

6.4.1. Wired Equivalent Privacy (WEP). WEP is an encryption *option* in the 802.11 standard required for WiFi™ certification (see Fig. 11).

WEP relies on the RC4 stream cipher [6] and a secret shared key. The secured payload consists of the data field and a 32-bit CRC (cyclic redundancy check), the integrity check value (ICV), of the data field, all of which are RC4 encrypted. The encryption seed consists of 40-bit shared secret key prepended with a 24-bit initialization vector (IV). The purpose of the IV is to randomize part of the key so that repeated encryptions of the same data are not

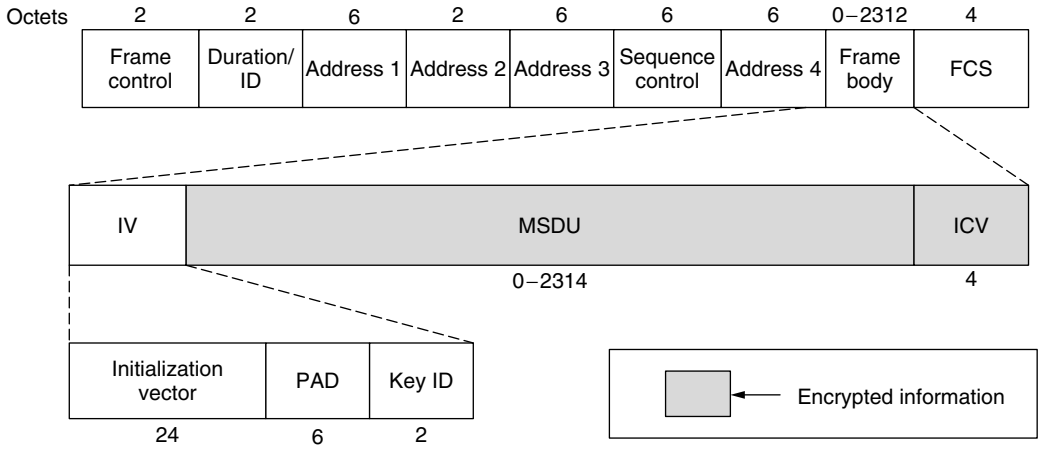


Figure 11. WAP frame body encryption.

identical, as this would provide a useful clue in breaking the encryption. The IV is transmitted in plaintext with the MAC frame. The rate at which the IV is changed is not specified in the standard but left to the discretion of the vendor.

6.4.2. Authentication Services. Two types of authentication services are specified within the 802.11 standard: *open system* and *shared key*. *Open-system* authentication automatically authenticates any station if the recipient station is operating in open system mode. The requesting station issues a message that contains its identity and a request for authentication. The recipient station responds with *successful* in the frame status code.

The *shared key* system differentiates stations that know a shared secret key from those that do not. The secret key is transmitted to the stations using an out-of-band mechanism and is stored as a write-only MIB (management information base) attribute. (Write-only attributes cannot be accessed *via* the MAC management system.)

Four frames are used in the authentication process. The first is the request for authentication by the station. The responding station sends the second frame containing a randomly generated 128-octet plaintext field to the requesting station. The requesting station encrypts the plaintext field and returns it to the responding station in the third frame. The responding station decrypts the message and compares the received plaintext to the original plaintext and, if they are identical, assumes that the requesting station knows the secret key and, therefore, should be authenticated. The responding station then sends the requesting station a successful authentication frame. If the decrypted text is not identical to the original plaintext message, then an unsuccessful authentication is returned.

6.4.3. Weaknesses of 802.11 Security. Almost concurrent with the development of the 802.11 standard were concerns for the strength of its security provisions. Researchers at the University of California at Berkeley [7] and others [8] have identified serious weaknesses in the WEP protocol. The concerns are centered on the way the RC4 cipher is employed rather than any fundamental weaknesses of the encryption technology itself. Four attacks have been identified.

The first attack, a *passive attack to decrypt traffic*, is based on the high potential volume of traffic coupled with the relatively short (24-bit) initialization vector. A system sending 1500-byte packets at 11 Mbps that is 10% utilized would be guaranteed to reuse the key in not more than 50 h. If the packets were smaller in size and the workload on the wireless system above 10%, a much shorter reuse period would occur. Further, these calculations assume that vendors change the IV for every transmission and use a random pattern with an externally generated seed. If multiple stations employ the same algorithm and start from the same seed, then reuse will occur much more quickly since all systems reinitialized will start from the same seed. Once multiple messages using the same IV have been recovered, statistical methods may be employed to identify the original contents of a message; all messages with that IV will be "open" for examination.

The second attack, an *active attack to inject traffic*, requires the plaintext of a message to be determined by the passive attack technique outlined above. Then using the relationship that RC4(X) xor X xor Y = RC4(Y), new messages can be encrypted successfully.

The third attack is an *active attack from both ends*. Here an attacker simply adjusts the bits of the header of a message changing the IP address to a host available to the attacker. This is possible since flipping a bit in the ciphertext flips the corresponding bit in the decrypted plaintext. It is also possible to calculate which bits of the CRC-32 must be flipped to compensate for the changes in the IP. All that is necessary is to capture a packet where the IP address is known and replay it on the system with the appropriate alternations. A plaintext IP packet will be sent to the controlled location on the Internet and examined to reveal the plaintext of the message. Then, any messages *using that same IV* can be decrypted.

The final attack is a *table-based attack* where a table of IVs and corresponding key streams are stored. Such a table would be large but eventually would allow a station to decrypt every packet sent through the system.

While these techniques require sophisticated monitoring and are not as simple as exploiting a hole in an operating system or application, they do imply that additional security work is required by the 802.11, Task Group i. WEP2 will employ 128-bit encryption and 128-bit IVs thus increasing the computation cost to break the system. The IEEE Task Group i has approved a draft to establish an authentication and key management system, tentatively called ESN *enhanced security network* (ESN), which will employ the draft Federal Information Processing Standard AES (advanced encryption standard).

6.5. WECA

The Wireless Ethernet Compatibility Alliance (WECA) [9] was formed in 1999 for the purpose of guaranteeing interoperability. Addressing 802.11, 2.4 GHz, DSSS, high data rate standards, this nonprofit organization's mission is to "certify interoperability of Wi-Fi (IEEE 802.11) products and to promote Wi-Fi as the global wireless LAN standard across all market segments." WECA has developed a certification test suite. Those products passing the compatibility tests are labeled Wi-Fi products.

6.6. Other Wireless Technologies

6.6.1. IEEE 802.11a. Standard 802.11a defines a PHY layer standard operating in the unlicensed 5 GHz band, known as U-NII (unlicensed national information infrastructure), and provides for data rates of 6, 9, 12, 18, 24, 36, 48, or 54 Mbps; 6, 12, and 24 Mbps are mandatory. It shares the same medium access controller (MAC) protocol as 802.11b. It achieves the higher data rate by using a higher carrier frequency along with orthogonal frequency-division multiplexing (OFDM) modulation.

OFDM survives multipath and intersymbol interference at these higher data rates by simultaneously transmitting multiple subcarriers on orthogonal frequency

channels where each subcarrier modulated at a low symbol rate.

A concern with 802.11a is that the lower power limit may restrict distances to less than 50 ft. Chipsets and products employing 802.11a are now (at the time of writing) beginning to appear but are considerably more expensive than those for 802.11b.

6.6.2. IEEE 802.11g. The IEEE 802.11 Task Group g is working toward a higher-speed version of 802.11b but has not yet approved a standard. To some, 802.11g appears to be in direct competition with 802.11a. For this and other reasons, 802.11g is politically charged, and its future is unclear at this time.

6.6.3. BlueTooth. BlueTooth [10] also operates in the 2.4 GHz ISM band using FHSS technology. Intended for personal area networking using low-cost interfaces for systems such as PDAs, mobile phones, and personal computers, it is limited to a 10-m distance. Bluetooth devices use the IEEE standard 48-bit addressing scheme. First generation devices operate up to 1 Mbps, while second-generation devices are expected to provide up to a 2 Mbps data rate.

More details on BlueTooth may be found elsewhere in the encyclopedia.

6.6.4. HomeRF. HomeRF [11] is a wireless network standard, specified in the shared wireless access protocol (SWAP) for home use where the distance requirements are limited. HomeRF 1.0 provides up to a 1.6 Mbps data rate at distances up to 150 ft. It uses FHSS transmission in the same 2.4 GHz band employed by 802.11b.

HomeRF incorporates the DECT (digital enhanced cordless telephony) standard to support mixed data and voice. The HomeRF 2.0 specification will support data rates in the 10 Mbps range. HomeRF also claims increased security over 802.11b. For example, it uses a 128-bit encryption key and a 32-bit IV that increases the IV reuse period significantly. Further, the way in which IVs are selected is specified in the protocol. Finally, the FHSS technology employed provides protection against denial of service attacks.

More details on HomeRF may be found elsewhere in the encyclopedia

7. THE FUTURE

Finally, the quest for increased speed continues. The 10 Gig (gigabit) Ethernet standard is in the final stages and expected to be completed and approved early in 2002, and 40 Gig Ethernet is being explored. The enormous popularity of Ethernet is not limited to *local* networks. Work is underway to employ variations of the Ethernet standard for metropolitan area networks currently employing PoS (packet over SONET) technology. Ethernet and its variations may eventually become the most popular standard for medium and long haul transmission. For example, the Metro Ethernet Forum [12] was created in June 2001 "to accelerate the adoption of optical Ethernet technology in metro networks around the globe."[1] The 10 Gig Ethernet contains a WAN PHY (physical) that describes a 40 km SMF (single-mode fiber) behavior suitable for competing with SONET. Additionally, the IEEE has created an 802.3ah *Ethernet in the First Mile* Task Force to explore the use of Ethernet fiber to the home/curb.

Although starting as a local network standard, Ethernet is rapidly expanding into the metropolitan and long haul environments and appears destined to become the prevalling standards for all classes of networking.

BIOGRAPHY

John H. Carson is a professor of management science at the George Washington University in Washington, D.C. Dr. Carson earned a B.S. in electrical engineering in 1969 and an M.S. and Ph.D. in information science in 1970 and 1976, respectively, from Lehigh University, Bethlehem, Pennsylvania. Dr. Carson served as manager of system engineering for the Software Productivity Consortium and as a principle scientist in the MITRE Corporation's networking center. He joined the George Washington University's Management Science Department in 1980, where he directed the M.S. program in information systems technology for 19 years. His areas of interest are software design, networking and communications technology, and information system design. Dr. Carson has authored and coauthored four books on computing and communications technology; he is a member of Eta Kapp Nu, Sigma Xi, and Beta Gamma Sigma.

BIBLIOGRAPHY

1. R. M. Metcalfe and D. R. Boggs, Ethernet: Distributed packet switching for local computer networks, *Commun. ACM* **19**(7): 395–404 (1976).
2. IEEE Std 802.3, 2000 ed., Part 3: *Carrier Sense Multiple Access with Collision Detection* (*CSMA/CD*) *Access Method and Physical Layer Specifications*, IEEE, 2000.
3. IEEE Standards for Local and Metropolitan Area Networks, *Overview and Architecture*, 1990.
4. IEEE 802.11 standards are available at *http://standards.ieee.org/getieee802/*.
5. B. O'Hara and A. Petrick, *802.11 Handbook, a Designer's Companion, IEEE Standards Information Network*, IEEE Press, 1999.
6. See papers on RC4 encryption from RSA Security at *http://www.rsa.com/*.
7. N. Borisov, I. Goldberg, and D. Wagner, Security of the WEP algorithm, *http://www.isaac.cs.berkeley.edu/isaac/wep-faq.html*, Univ. California at Berkeley, Feb. 2001.
8. P. C. Mehta, Wired equivalent privacy vulnerability, *http://www.sans.org/infosecFAQ/wireless/equiv.htm*, SANS Institute, April 4, 2001.
9. *http://www.weca.net*.
10. *http://www.bluetooth.com*.
11. *http://www.homerf.org*.
12. *www.metroethernetforum.org/*.

LOOP ANTENNAS

KAZIMIERZ (KAI) SIWIAK
Time Domain Corporation
Huntsville, Alabama

1. INTRODUCTION

The *IEEE Standard Definitions of Terms for Antennas* [1] defines the loop as "an antenna whose configuration is that of a loop," further noting that "if the current in the loop, or in the multiple parallel turns of the loop, is essentially uniform and the loop circumference is small compared with the wavelength, the radiation pattern approximates that of a magnetic dipole." That definition and the further note imply the two basic realms of loop antennas: electrically small, and electrically large structures.

There are hundreds of millions of loop antennas currently in use [2] by subscribers of personal communications devices, primarily pagers. Furthermore, loops have appeared as transmitting arrays, like the massive multielement loop array at shortwave station HCJB in Quito, Ecuador, and as fractional wavelength-size tunable HF transmitting antennas. The loop is indeed an important and pervasive communications antenna.

The following analysis of loop antennas reveals that the loop, when small compared with a wavelength, exhibits a radiation resistance proportional to the square of the enclosed area. Extremely low values of radiation resistance are encountered for such loops, and extreme care must be taken to effect efficient antenna designs. Furthermore, when the small loop is implemented as a transmitting resonant circuit, surprisingly high voltages can exist across the resonating capacitor even for modest applied transmitter power levels. The wave impedance in the immediate vicinity of the loop is low, but at close distances (0.1–2 wavelengths) exceeds the intrinsic free space impedance before approaching that value.

A loop analysis is summarized that applies to loops of arbitrary circular diameter and of arbitrary wire thickness. The analysis leads to some detail regarding the current density in the cross section of the wire. Loops of shapes other than circular are less easily analyzed, and are best handled by numerical methods such as moment method described by Burke and Poggio [3].

Loops are the antennas of choice in pager receivers, and appear as both ferrite loaded loops and as single-turn rectangular shaped structures within the radio housing. Body worn loops benefit from a field enhancement because of the resonant behavior of human body with respect to vertically polarized waves. In the high frequency bands, the loop is used as a series resonant circuit fed by a secondary loop. The structure can be tuned over a very large frequency band while maintaining a relatively constant feed point impedance. Large loop arrays comprised of one wavelength perimeter square loops have been successfully implemented as high-gain transmitting structures at high power shortwave stations.

2. ANALYSIS OF LOOP ANTENNAS

Loop antennas, particularly circular loops, were among the first radiating structures analyzed beginning as early as 1897 with Pocklington's analysis [4] of the thin wire loop excited by a plane wave. Later, Hallén [5] and Storer [6] studied driven loops. All these authors used a Fourier expansion of the loop current, and the latter two authors discovered numerical difficulties with the approach. The difficulties could be avoided, as pointed out by Wu [7], by integrating the Green function over the toroidal surface of the surface of the wire. The present author coauthored an improved theory [8,9] that specifically takes into account the finite dimension of the loop wire and extends the validity of the solution to fatter wires than previously considered. Additionally, the work revealed some detail of the loop current around the loop cross section. Arbitrarily shaped loops, such as triangular loops and square loops, as well as loop arrays can be conveniently analyzed using numerical methods.

2.1. The Infinitesimal Loop Antenna

The infinitesimal current loop consists of a circulating current I enclosing an infinitesimal surface area S, and is solved by analogy to the infinitesimal dipole. The fields of an elementary loop element of radius b can be written in terms of the loop enclosed area $S = \pi b^2$ and a constant excitation current, I (when I is RMS, then the fields are also RMS quantities). The fields are "near" in the sense that the distance parameter r is far smaller than the wavelength, but far larger than the loop dimension $2b$. Hence, this is *not* the *close* near field region. The term kIS is often called the *loop moment* and is analogous to the similar term Ih associated with the *dipole moment*. The infinitesimally small loop is pictured in Fig. 1a next to its elementary dipole analog (Fig. 1b). The dipole uniform current I flowing over an elemental length h is the dual of a "magnetic current" $M_z S = Ih$ and the surface area is $S = h/k$. The fields due to the infinitesimal loop are then found from the vector and scalar potentials.

2.1.1. Vector and Scalar Potentials. The wave equation, in the form of the inhomogeneous Helmholtz equation, is used here with most of the underlying vector arithmetic omitted; see Refs. 10–12 for more details. For a magnetic current element source, the electric displacement **D** is always solenoidal (the field lines do not originate or

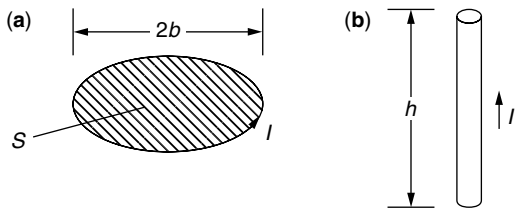


Figure 1. Small-antenna geometry showing (**a**) the parameters of the infinitesimal loop moment, and (**b**) its elementary dipole dual. (*Source*: Siwiak [2].)

terminate on sources); that is, in the absence of source charges the divergence is zero

$$\nabla \cdot \mathbf{D} = 0 \tag{1}$$

and the electric displacement field can be represented by the curl of an arbitrary vector $\mathbf{F}$

$$\mathbf{D} = \varepsilon_0 \mathbf{E} = \nabla \times \mathbf{F} \tag{2}$$

where $\mathbf{F}$ is the vector potential and obeys the vector identity $\nabla \cdot \nabla \times \mathbf{F} = 0$. Using Ampere's law in the absence of electric sources, we obtain

$$\nabla \times \mathbf{H} = j\omega\varepsilon_0 \mathbf{E} \tag{3}$$

and with the vector identity $\nabla \times (-\nabla\Phi) = 0$, where Φ represents an arbitrary scalar function of position, it follows that

$$\mathbf{H} = -\nabla\Phi - j\omega\mathbf{F} \tag{4}$$

and for a homogeneous medium, after some manipulation, we get

$$\nabla^2\mathbf{F} + k^2\mathbf{F} = -\varepsilon_0\mathbf{M} + \nabla(\nabla \cdot \mathbf{F} + j\omega\mu_0\varepsilon_0\Phi) \tag{5}$$

where k is the wavenumber and $k^2 = \omega^2\mu_0\varepsilon_0$. Although equation (2) defines the curl of $\mathbf{F}$, the divergence of $\mathbf{F}$ can be independently defined and the *Lorentz condition* is chosen

$$j\omega\mu_0\varepsilon_0\Phi = -\nabla \cdot \mathbf{F} \tag{6}$$

where ∇^2 is the Laplacian operator given by

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{7}$$

Substituting the simplification of Eq. (6) into (5) leads to the inhomogeneous Helmholtz equation

$$\nabla^2\mathbf{F} + k^2\mathbf{F} = -\varepsilon_0\mathbf{M} \tag{8}$$

Similarly, by using Eqs. (6) and (4) it is seen that

$$\nabla^2\Phi + k^2\Phi = 0 \tag{9}$$

Using Eq. (4) and the Lorentz condition of Eq. (6), we can find the electric field solely in terms of the vector potential $\mathbf{F}$. The utility of that definition becomes apparent when we consider a magnetic current source aligned along a single vector direction, for example, $\mathbf{M} = \mathbf{z}M_z$, for which the vector potential is $\mathbf{F} = \mathbf{z}F_z$, where $\mathbf{z}$ is the unit vector aligned with the z axis, and Eq. (8) becomes a scalar equation.

2.1.2. Radiation from a Magnetic Current Element. The solution to the wave equation (8) presented here, with the details suppressed, is a spherical wave. The results are used to derive the radiation properties of the infinitesimal current loop as the dual of the infinitesimal current element. The infinitesimal magnetic current element $\mathbf{M} = \mathbf{z}M_z$ located at the origin satisfies a one-dimensional, and hence scalar form of Eq. (8). At points excluding the origin where the infinitesimal current element is located, Eq. (8) is source-free and is written as a function of radial distance r

$$\nabla^2 F_z(r) + k^2 F_z(r) = \frac{1}{r^2}\frac{\partial}{\partial r}\left[r^2\frac{\partial F_z(r)}{\partial r}\right] + k^2 F_z(r) = 0 \tag{10}$$

which can be reduced to

$$\frac{d^2F_z(r)}{dr^2} + \frac{2}{r}\frac{dF_z(r)}{dr} + k^2F_z(r) = 0 \tag{11}$$

Since F_z is a function of only the radial coordinate, the partial derivative in Eq. (10) was replaced with the ordinary derivative. Eq. (11) has a solution

$$F_z = C_1\frac{e^{-jkr}}{r} \tag{12}$$

There is a second solution where the exponent of the phasor quantity is positive; however, we are interested here in outward traveling waves, so we discard that solution. In the static case the phasor quantity is unity. The constant C_1 is related to the strength of the source current, and is found by integrating Eq. (8) over the volume, including the source giving

$$C_1 = \frac{\varepsilon_0}{4\pi}kIS \tag{13}$$

and the solution for the vector potential is in the $\mathbf{z}$ unit vector direction

$$\mathbf{F} = \frac{\varepsilon_0}{4\pi}kIS\frac{e^{-jkr}}{r}\mathbf{z} \tag{14}$$

which is an outward propagating spherical wave with increasing phase delay (increasingly negative phase) and with amplitude decreasing as the inverse of distance. We may now solve for the magnetic fields of an infinitesimal current element by inserting Eq. (14) into (4) with Eq. (6) and then for the electric field by using Eq. (2). The fields, after sufficient manipulation, and for $r \gg kS$, are

$$H_r = \frac{kIS}{2\pi}e^{-jkr}k^2\left[\frac{j}{(kr)^2} + \frac{1}{(kr)^3}\right]\cos(\theta) \tag{15}$$

$$H_\theta = \frac{kIS}{4\pi}e^{-jkr}k^2\left[-\frac{1}{kr} + \frac{j}{(kr)^2} + \frac{1}{(kr)^3}\right]\sin(\theta) \tag{16}$$

$$E_\phi = \eta_0\frac{kIS}{4\pi}e^{-jkr}k^2\left[\frac{1}{kr} - \frac{j}{(kr)^2}\right]\sin(\theta) \tag{17}$$

where $\eta_0 = c\mu_0 = 376.730313$ is the intrinsic free-space impedance, c is the velocity of propagation (see Ref. 13 for definitions of constants), and I is the loop current.

Equations (15) and (16) for the *magnetic* fields H_r and H_θ (1.30) of the infinitesimal loop have exactly the same form as the *electric* fields E_r and E_θ for the infinitesimal dipole, while Eq. (17) for the *electric* field of the loop E_ϕ has exactly the same form as the *magnetic* field H_ϕ of the dipole when the term kIS of the loop expressions is replaced with Ih for the infinitesimal ideal (uniform current element) dipole. In the case where the loop moment

kIS is superimposed on, and equals the dipole moment Ih, the fields in all space will be circularly polarized.

Equations (15)–(17) describe a particularly complex field behavior for what is a very idealized selection of sources: a simple linear magnetic current M representing a current loop I encompassing an infinitesimal surface $S = \pi b^2$. Expressions (15)–(17) are valid only in the region sufficiently far ($r \gg kS$) from the region of the magnetic current source M.

2.1.3. The Wave Impedance of Loop Radiation. The wave impedance can be defined as the ratio of the total electric field magnitude divided by the total magnetic field magnitude. We can study the wave impedance of the loop fields by using Eqs. (15)–(17) for the infinitesimal loop fields, along with their dual quantities for the ideal electric dipole. Figure 2 shows the loop field wave impedance as a function of distance kr from the loop along the direction of maximum far field radiation. The wave impedance for the elementary dipole is shown for comparison. At distances near $kr = 1$ the wave impedance of loop radiation exceeds $\eta_0 = 376.73\ \Omega$, the intrinsic free-space impedance, while that of the infinitesimal loop is below 376.73 Ω. In this region, the electric fields of the loop dominate.

2.1.4. The Radiation Regions of Loops. Inspection of Eqs. (15)–(17) for the loop reveal a very complex field structure. There are components of the fields that vary as the inverse third power of distance r, inverse square of r, and the inverse of r. In the near field or induction region of the idealized infinitesimal loop, that is, for $kr \ll 1$ (however, $r \gg kS$ for the loop and $r \gg h$ for the dipole), the magnetic fields vary as the inverse third power of distance.

The region where kr is nearly unity is part of the radiating near field of the Fresnel zone. The inner boundary of that zone is taken by Jordan [12] to be

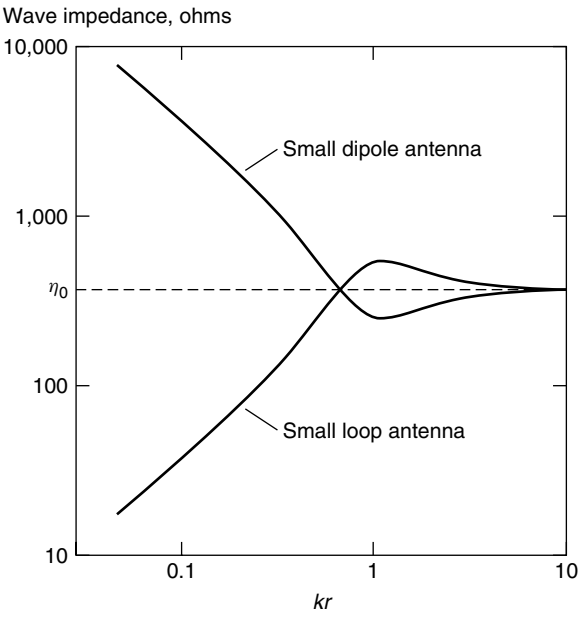


Figure 2. Small loop antenna and dipole antenna wave impedances compared. (*Source*: Siwiak [2].)

$r^2 > 0.38D^3/\lambda$, and the outer boundary is $r < 2D^2/\lambda$, where D is the largest dimension of the antenna, here equal to $2b$. The outer boundary criterion is based on a maximum phase error of $\pi/8$. There is a significant radial component of the field in the Fresnel zone.

The far-field or Fraunhofer zone is a region of the field for which the angular radiation pattern is essentially independent of distance. That region is usually defined as extending from $r < 2D^2/\lambda$ to infinity, and the field amplitudes there are essentially proportional to the inverse of distance from the source. Far-zone behavior is identified with the basic free space propagation law.

2.1.5. The Induction Zone of Loops. We can study the "induction zone" in comparison to the "far field" by considering "induction zone" coupling that was investigated by Hazeltine [14] and that was applied to low frequency radio receiver designs of his time. Today the problem might be applied to the design of a miniature radio module where inductors must be oriented for minimum coupling. The problem Hazeltine solved was one of finding the geometric orientation for which two loops in parallel planes have minimum coupling in the induction zone of their near fields and serves to illustrate that the "near field" behavior differs fundamentally and significantly from "far field" behavior. To study the problem we invoke the principle of reciprocity, which states

$$\int_V \left[\mathbf{E}_b \cdot \mathbf{J}_a - \mathbf{H}_b \cdot \mathbf{M}_a\right] dV \equiv \int_V \left[\mathbf{E}_a \cdot \mathbf{J}_b - \mathbf{H}_a \cdot \mathbf{M}_b\right] dV \tag{18}$$

That is, the reaction on antenna (a) of sources (b) equals the reaction on antenna (b) of sources (a). For two loops with loop moments parallel to the z axis, we want to find the angle θ for which the coupling between the loops vanishes; that is, both sides of equation (18) are zero. The reference geometry is shown in Fig. 3. In the case of the loop, there are no electric sources in Eq. (18), so $\mathbf{J}_a = \mathbf{J}_b = 0$, and both $\mathbf{M}_a$ and $\mathbf{M}_b$ are aligned with $\mathbf{z}$, the unit vector parallel to the z axis. Retaining only the inductive field components and clearing common constants in Eqs. (15) and (17) are placed into (18). We require that $(H_r\mathbf{r} + H_\theta\theta)\mathbf{z} = 0$. Since $\mathbf{r} \cdot \mathbf{z} = -\sin(\theta)$ and $\mathbf{r} \cdot \mathbf{z} = \cos(\theta)$, we are left with $2\cos^2(\theta) - \sin^2(\theta) = 0$, for which $\theta = 54.736°$. When oriented as shown in Fig. 3, two loops parallel to the x–y plane whose centers are displaced by an angle of 54.736° with respect to the z axis will not couple in their near fields. To be sure, the angle determined above is "exactly" correct for infinitesimally small loops; however, that angle will be nominally the same for larger loops. Hazeltine [14] used this principle, placing the axes of the inductors in a common plane each at an angle of 54.7° with respect to the normal form the radio chassis, to minimize the coupling between the inductors.

The same principle can be exploited in the design of a metal detector, as depicted in Fig. 4. The loop a is driven with an audiofrequency oscillator. Loop b, in a parallel plane and displaced so that nominally $\theta = 54.7°$, is connected to a detector that might contain an audio amplifier that feeds a set of headphones. Any conductive object near loop a will disrupt the balance of the system

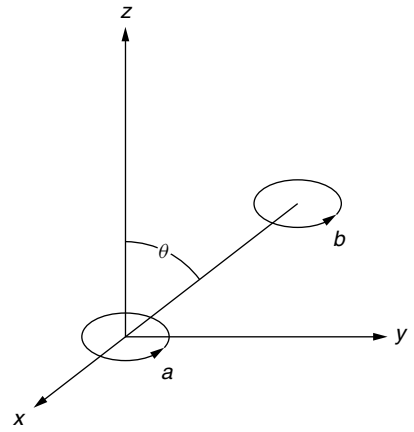


Figure 3. Two small loops in parallel planes and with $\theta = 54.736^\circ$ will not couple in their near fields. (*Source*: Siwiak [2].)

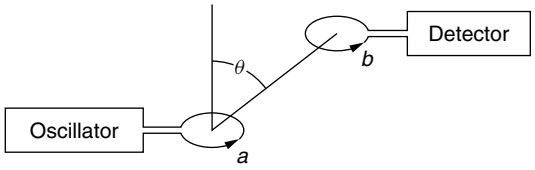


Figure 4. A metal detector employs two loops initially oriented to minimize coupling in their near fields.

and result in an increased coupling between the two loops, thus indicating the presence of a conducting object near a.

2.1.6. The Intermediate- and Far-Field Zones of Loops. The loop coupling problem provides us with a way to investigate the intermediate and far-field coupling by applying Eq. (18) with Eqs. (15) and (16) for various loop separations kr. In the far-field region only the H_θ term of the magnetic field survives, and by inspection of Eq. (16), the minimum coupling occurs for $\theta = 0$ or 180°. Figure 5 compares the coupling (normalized to their peak values) for loops in parallel planes whose fields are given by Eq. (15)–(17). Figure 5 shows the coupling as a function of angle θ for an intermediate region ($kr = 2$) and for the far-field case ($kr = 1000$) in comparison with the induction-zone case ($kr = 0.001$). The patterns are fundamentally and significantly different. The coupling null at $\theta = 54.7^\circ$ is clearly evident for the induction-zone case $kr = 0.001$ and for which the $(1/kr)^3$ terms dominate. Equally evident is the far-field coupling null for parallel loops on a common axis when the $1/kr$ terms dominate. The intermediate zone coupling shows a transitional behavior where all the terms in kr are comparable.

2.1.7. The Directivity and Impedance of Small Loops. The *directive gain* of the small loop can be found from the far-field radially directed Poynting vector in ratio to the average Poynting vector over the radian sphere:

$$D(\theta,\phi) = \frac{|(\mathbf{E}\times\mathbf{H})\cdot\mathbf{r}|}{\dfrac{1}{4\pi}\displaystyle\int_0^{2\pi}\int_0^{\pi}|(\mathbf{E}\times\mathbf{H})\cdot\mathbf{r}|\sin(\theta)\,d\theta\,d\phi} \tag{19}$$

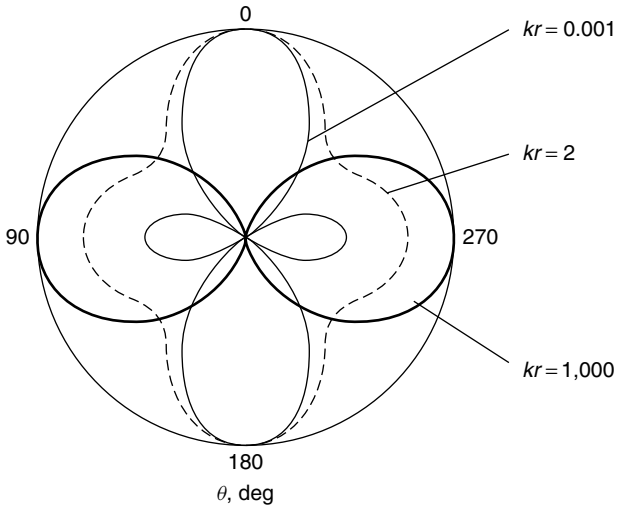


Figure 5. Normalized-induction-zone, intermediate-zone, and far-zone coupling between loops in parallel planes. (*Source*: Siwiak [2].)

Only the θ component of H and the ϕ component of E survive into the far field. Using Eq. (16) for H_θ and Eq. (17) for E_ϕ and retaining only the $1/kr$ terms, we see that Eq. (19) yields $D = 1.5\sin^2(\theta)$ by noting that the functional form of the product of E and H is simply $\sin^2(\theta)$ and by carrying out the simple integration in the denominator of Eq. (19).

Taking into account the directive gain, the far-field power density P_d in the peak of the pattern is

$$P_\mathrm{density} = \frac{1.5\,I^2R_\mathrm{radiation}}{4\pi r^2} = H_\theta^2\eta_0 = \left[\frac{kS}{4\pi}\frac{k}{r}I\right]^2\eta_0 \tag{20}$$

For radiated power $I^2R_\mathrm{radiation}$, hence, we can solve for the radiation resistance

$$R_\mathrm{radiation} = \frac{(kS)^2}{6\pi}\eta_0 = \eta_0\frac{\pi}{6}(kb)^4 \tag{21}$$

for the infinitesimal loop of loop radius b.

When fed by a gap, there is a dipole moment that adds terms not only to the impedance of the loop but also to the close near fields. For the geometry shown in Fig. 6, and using the analysis of King [15], the electrically small loop, having a diameter $2b$ and wire diameter $2a$, exhibits a feed point impedance given by

$$Z_\mathrm{loop} = \eta_0\frac{\pi}{6}(kb)^4\left[1+8(kb)^2\right]\left[1-\frac{a^2}{b^2}\right]+\cdots$$
$$+\,j\eta_0 kb\left[\ln\left[\frac{8b}{a}\right]-2+\frac{2}{3}(kb)^2\right][1+2(kb)^2] \tag{22}$$

including dipole-mode terms valid for $kb \ll 0.1$. The leading term of Eq. (22) is the same as derived in Eq. (21) for the infinitesimal loop. Expression (22) adds the detail of terms considering the dipole moment of the gap-fed loop as well as refinements for loop wire radius a. The small-loop antenna is characterized by a radiation resistance that is proportional to the Fourth power of the loop radius b.

The reactance is inductive; hence, it is proportional to the antenna radius. It follows that the Q is inversely proportional to the third power of the loop radius, a result that is consistent with the fundamental limit behavior for small antennas.

Using Eq. (22), and ignoring the dipole-mode terms and second order terms in a/b, the unloaded Q of the loop antenna, is

$$Q_{\text{loop}} = \frac{\frac{6}{\pi}\left[\ln\left[\frac{8b}{a}\right] - 2\right]}{(kb)^3} \tag{23}$$

which for $b/a = 6$ becomes

$$Q_{\text{loop}} = \frac{3.6}{(kb)^3} \tag{24}$$

which has the proper limiting behavior for small-loop radius. The Q of the small loop given by Eq. (23) is indeed larger than the minimum possible $Q_{\min} = (kb)^{-3}$ predicted in Siwiak [2] for a structure of its size. It must be emphasized that the actual Q of such an antenna will be smaller than given by Eq. (24) because of unavoidable dissipative losses not represented in Eq. (22)–(24). We can approach the minimum Q but never be smaller, except by introducing dissipative losses.

2.2. The Gap-Fed Loop

The analysis of arbitrarily fat wire loops follows the method in Ref. 8, shown in simplified form in Ref. 9 and summarized here. The toroid geometry of the loop is expressed in cylindrical coordinates ρ, ϕ, and z with the toroid located symmetrically in the $z = 0$ plane. The relevant geometry is shown in Fig. 6.

2.2.1. Loop Surface Current Density. The current density on the surface of the toroidal surface of the loop is given by

$$J_\phi = \sum_{n=-\infty}^{\infty} \sum_{p=-\infty}^{\infty} A_{n,p} e^{jn\phi} F_p \tag{25}$$

where the functions F_p are symmetric about the z axis and are simple functions of $\cos(n\psi)$, where ψ is in the cross section of the wire as shown in Fig. 6 and is related to the cylindrical coordinate by $z = a\sin(\psi)$. These function

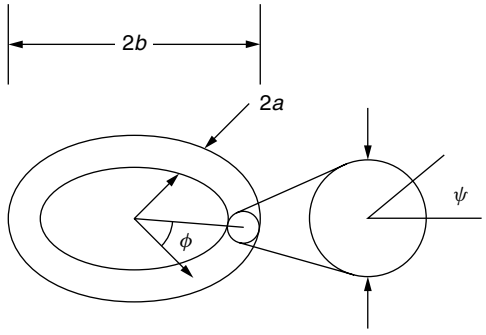


Figure 6. Parameters of the fat wire loop. (*Source*: Siwiak [2].)

are orthonormalized over the conductor surface using the Gram–Schmidt method described in (16), yielding

$$F_0 = \frac{1}{2\pi\sqrt{ab}} \tag{26}$$

and

$$F_1 = F_0\sqrt{\frac{2}{1-(a/2b)^2}}\left[\cos(\psi) - \frac{a}{2b}\right] \tag{27}$$

The higher-order functions are lengthy but simple functions of $\sin(p\psi)$ and $\cos(p\psi)$.

2.2.2. Scalar and Vector Potentials. The electric field is obtained form the vector and scalar potentials

$$\mathbf{E} = -\nabla\Phi - j\omega\mathbf{A} \tag{28}$$

The boundary conditions require that E_ϕ, E_ψ, and E_ρ are zero on the surface of the loop everywhere except at the feed gap $|\phi| \le \varepsilon$. Because this analysis will be limited to wire diameters significantly smaller than a wavelength, the boundary conditions on E_ψ and E_ρ will not be enforced. In the gap $E_\phi = V_0/2\varepsilon\rho$, where V_0 is the gap excitation voltage.

The components of the vector potential are simply

$$A_\phi = \frac{1}{4\pi}\int_S\int J_\phi \cos(\phi - \phi')\, dS \tag{29}$$

and

$$A_\rho = \frac{1}{4\pi}\int_S\int J_\phi \sin(\phi - \phi')\, dS \tag{30}$$

and the vector potential is

$$\Phi = \frac{j\eta_0}{4\pi k}\int_S\int \frac{1}{\rho}\frac{\partial J_\phi}{\partial\phi} G\, dS \tag{31}$$

where the value of $dS = [b + a\sin(\psi)]a\, d\psi$. *Green's function* G is expressed in terms of cylindrical waves to match the rotational symmetry of the loop

$$G = \frac{1}{2j}\sum_{m=-\infty}^{\infty} e^{-jm(\phi-\phi')}\int_{-\infty}^{\infty} J_m(\rho_1 - \nu)H_m^{(2)}(\rho_2 - \nu)e^{-j\zeta(z-z')}\, d\zeta \tag{32}$$

where $\nu = \sqrt{k^2 + \zeta^2}$
$\rho_1 = \rho - a\cos(\psi)$
$\rho_2 = \rho + a\cos(\psi)$

and where $J_m(\nu\rho)$ and $H_m^{(2)}(\nu\rho)$ are the Bessel and Hankel functions, respectively.

2.2.3. Matching the Boundary Conditions. Expression (8) is now inserted into Eqs. (5)–(8), and the electric field is then found from Eq. (2) and the boundary condition is enforced. For constant ρ on the wire

$$\int_{-\pi}^{\pi} E_\phi e^{jm\phi}\, d\phi = -\frac{V_0}{\rho}\frac{\sin(m\varepsilon)}{m\varepsilon} \tag{33}$$

This condition is enforced on the wire as many times as there are harmonics in ψ. Truncating the index p as

described in Ref. 9 to a small finite number P, we force $E_\phi = 0$ except in the feeding gap along the lines of constant ρ on the surface of the toroid. If we truncate to P, the number of harmonics F_p in ψ and to M the number of harmonics in ϕ, we find the radiation current by solving M systems of $P \times P$ algebraic equations in $A_{m,p}$. In Ref. 9, $P = 2$ and M in the several hundreds was found to be a reasonable computational task that led to useful solutions.

2.2.4. Loop Fields and Impedance. With the harmonic amplitudes $A_{m,p}$ known, the current density is found from Eq. (1). The electric field is found next from Eq. (2) and the magnetic field is given by

$$H_\rho = -\frac{\partial A_\phi}{\partial z} \tag{34}$$

$$H_\phi = -\frac{\partial A_\rho}{\partial z} \tag{35}$$

$$H_z = \frac{\partial A_\phi}{\partial \rho} + \frac{A_\phi}{\rho} - \frac{1}{\rho}\frac{\partial A_\rho}{\partial \rho} \tag{36}$$

The loop current across a section of the wire is found by integrating the function J_ϕ in Eq. (25) around the wire cross section. The loop radiation impedance is then the applied voltage V_0 in the gap divided by the current in the gap. Figure 7 shows the loop feed radiation resistance, and Fig. 8 shows the corresponding loop reactance, as a function of loop radius kr for a thin wire, $\Omega = 15$, and a fat wire, $\Omega = 10$, loop where $\Omega = 2\ln(2\pi b/a)$. The thin-wire loop has very sharp resonant behavior compared with the fat-wire loop, especially for a half-wavelength-diameter ($kb = 0.5$) structure. The higher resonances are less pronounced for both loops. Fat-wire loops exhibit an interesting behavior in that at a diameter of about a half-wavelength, the reactance is essentially always capacitive and the total impedance remains well behaved.

2.2.5. Small Gap-Fed Loops. The detailed analysis of the fat, gap-fed wire loop, as shown in Refs. 8 and 9, reveals

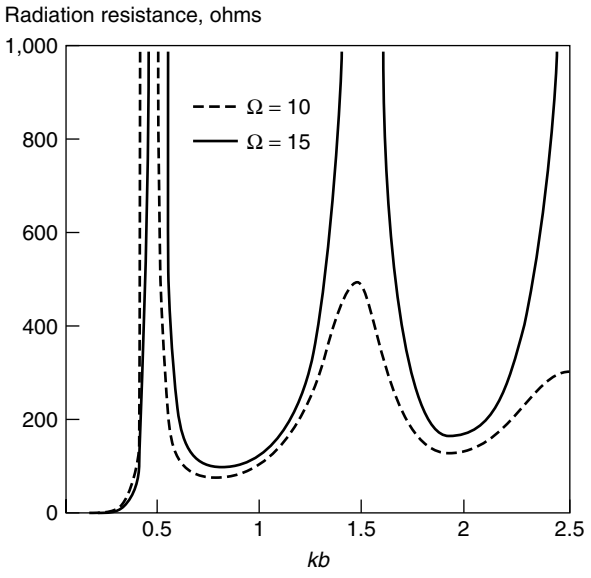


Figure 7. Loop radiation resistance.

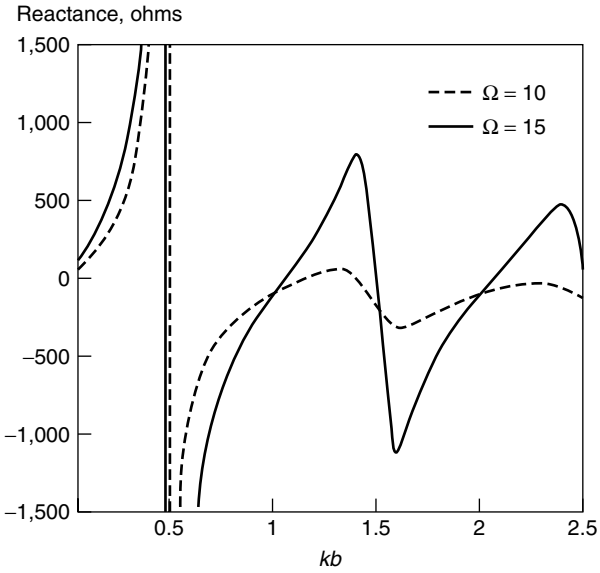


Figure 8. Loop reactance.

that the current density around the circumference of the wire, angle ψ in Fig. 6, is not constant. An approximation to the current density along the wire circumference for a small-diameter loop is

$$J_\phi = \frac{I_\phi}{2\pi a}[1 - 2\cos(\phi)(kb)^2][1 + Y\cos(\psi)] \tag{37}$$

where I_ϕ is the loop current, which has cosine variation along the *loop circumference*; and where the variation around the *wire circumference* is shown as a function of the angle ψ. Y is the ratio of the first to the zero-order mode in ϕ, and is not a simple function of loop dimensions a and b, but can be found numerically in Siwiak [2] and from the analysis of the previous section. For the small loop Y is negative and of order a/b, so Eq. (37) predicts that there is current bunching along the inner contour ($\psi = 180^\circ$) of the wire loop. Table 1 gives representative values for Y as a function of a/b.

This increased current density results in a corresponding increase in dissipative losses in the small loop. We can infer that the cross-sectional shape of the conductor formed into a loop antenna will impact the loss performance in a small loop.

The small loop fed with a voltage gap has a charge accumulation at the gap and will exhibit a close near electric field. For a small loop of radius b and centered in

Table 1. Parameter Y for Various Loop Thickness and $b = 0.01$ Wavelengths

Ω	a/λ	Y
19.899	0.000003	−0.0039
17.491	0.00001	−0.0090
15.294	0.00003	−0.020
12.886	0.0001	−0.048
10.689	0.0003	−0.098
8.2809	0.001	−0.179

the x–y plane, the fields at $(x, y) = (0, 0)$ are derived in Ref. 9 and given here as

$$E_\phi = -j\frac{\eta_0 kI}{2} \tag{38}$$

where I is the loop current and

$$H_z = \frac{I}{2b} \tag{39}$$

Expression (39) is recognized as the classic expression for the static magnetic field within a single-turn solenoid. Note that the electric field given by Eq. (38) does not depend on any loop dimensions, but was derived for an electrically small loop. The wave impedance, Z_w, at the origin, is the ratio of E_ϕ to H_z and from Eqs. (38) and (39) is

$$Z_w = -j\eta_0 kb \tag{40}$$

In addition to providing insight into the behavior of loop probes, Eqs. (38)–(40) are useful in testing the results of numerical codes like the numerical electromagnetic code (NEC) described in Ref. 3, and often used in the numerical analysis of wire antenna structures.

When the small loop is used as an untuned and unshielded field probe, the current induced in the loop will have a component due to the magnetic field normal to the loop plane as well as a component due to the electric field in the plane of the loop. A measure of E field to H field sensitivity is apparent from expression (40). The electric field to magnetic field sensitivity of a simple small-loop probe is proportional to the loop diameter. The small gap-fed loop, then, has a dipole moment that complicates its use as a purely magnetic field probe.

3. LOOP APPLICATIONS

Loop antennas appear in pager receivers as both ferrite-loaded loops and as single-turn rectangular structure within the radio housing. When worn on the belt, the loop benefits from coupling to the vertically resonant human body. In the high-frequency bands, the loop has been implemented as a series resonant circuit fed by a secondary loop. The structure can be tuned over a very large frequency band while maintaining a relatively constant feed point impedance. One wavelength perimeter square loops have been successfully implemented as high-gain transmitting structures.

3.1. The Ferrite-Loaded Loop Antenna: A Magnetic Dipole

Let us examine a small ferrite-loaded loop antenna with dimensions, $2h = 2.4$ cm, $2a = 0.4$ cm, and at a wavelength of about $\lambda = 8.6$ m as depicted in Fig. 9. When the permeability of the ferrite is sufficiently high, this antenna behaves like a magnetic dipole. The magnetic fields are strongly confined to the magnetic medium, especially near the midsection of the ferrite rod, and behave as the dual of the electric dipole excited by a triangular current distribution. We can therefore analyze its behavior using a small dipole analysis shown in Siwiak [2].

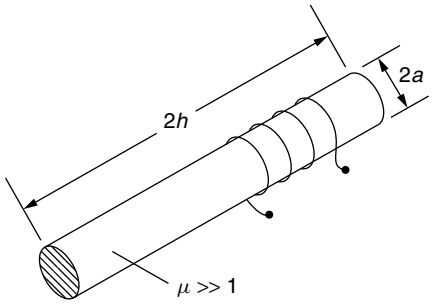


Figure 9. A ferrite loaded loop antenna. (*Source*: Siwiak [2].)

The impedance at the midpoint of a short dipole having a current uniformly decreasing from the feed point across its length $2h$ is

$$Z_{\text{dipole}} = \frac{\eta_0}{6\pi}(kh)^2 - j\frac{\dfrac{\eta_0}{2\pi}\left[\ln\left[\dfrac{2h}{a}\right] - 1\right]}{kh} \tag{41}$$

The corresponding unloaded Q of the dipole antenna is

$$Q_{\text{dipole}} = \frac{3\left[\ln\left[\dfrac{2h}{a}\right] - 1\right]}{(kh)^3} \tag{42}$$

Equation (42) has the expected inverse third power with size behavior for small antennas, and for $h/a = 6$

$$Q_{\text{dipole}} = \frac{4.5}{(kh)^3} \tag{43}$$

Comparing the Q for a small dipole given by Eq. (43) with the Q of a small loop of Eq. (24), we see that the loop Q is small even though the same ratio of antenna dimension to wire radius was used. We conclude that the small loop utilizes the smallest sphere that encloses it more efficiently than does the small dipole. Indeed, the thin dipole is essentially a one-dimensional structure, while the small loop is essentially a two-dimensional structure.

We can use Eqs. (41) and (42) for the elementary dipole to examine the ferrite load loop antenna since it resembles a magnetic dipole. The minimum ideal Q of this antenna is given by Eq. (42), 1.0×10^6. The corresponding bandwidth of such an antenna having no dissipative losses would be $2 \times 35f/Q = 70$ MHz$/1.3 \times 10^6 = 69$ Hz. A practical ferrite antenna at this frequency has an actual unloaded Q_A of nearer to 100, as can be inferred from the performance of belt-mounted radios shown in Table 2. Hence, an estimate of the actual antenna efficiency is

$$10\log\frac{Q_A}{Q} = -40\,\text{dB} \tag{44}$$

and the actual resultant 3 dB bandwidth is about 700 kHz. Such an antenna is typical of the type that would be used in a body-mounted paging receiver application. As detailed in Siwiak [2], the body exhibits an average magnetic field

Table 2. Paging Receiver Performance Using Loops

Frequency Band (MHz)	Paging Receiver, at Belt Average Gain (dBi)	Field Strength Sensitivity (dB·μV/m)
30–50	−32 to −37	12–17
85	−26	13
160	−19 to −23	10–14
280–300	−16	10
460	−12	12
800–960	−9	18–28

Source: After Siwiak [2].

enhancement of about 6 dB at this frequency, so the average belt-mounted antenna gain is −34 dBi. This is typical of a front-position body-mounted paging or personal communication receiver performance in this frequency range.

3.2. Body Enhancement in Body-Worn Loops

Loops are often implemented as internal antennas in pager receiver applications spanning the frequency bands from 30 to 960 MHz. Pagers are often worn at belt level, and benefit from the "body enhancement" effect. The standing adult human body resembles a lossy wire antenna that resonates in the range of 40–80 MHz. The frequency response, as seen in Fig. 10, is broad, and for belt-mounted loop antennas polarized in the body axis direction, enhances the loop antenna azimuth-averaged gain at frequencies below about 500 MHz.

The far-field radiation pattern of a body-worn receiver is nearly omnidirectional at very low frequency. As frequency is increased, the pattern behind the body develops a shadow that is manifest as a deepening null with increasing frequency. In the high-frequency limit, there is only a forward lobe with the back half-space essentially completely blocked by the body. For horizontal incident polarization, there is no longitudinal body resonance and there is only slight enhancement above 100 MHz.

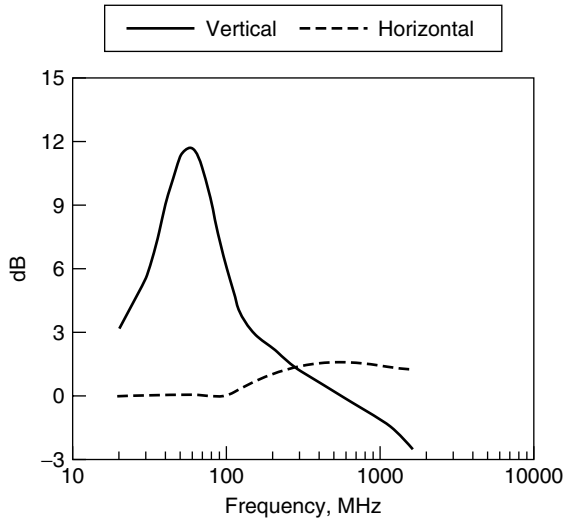


Figure 10. Gain-averaged body-enhanced loop response. (*Source*: Siwiak [2].)

3.3. The Small Resonated High-Frequency Loop

The simple loop may be resonated with a series capacitor having a magnitude of reactance equal to the loop reactance, and indeed, is effectively implemented that way for use in the HF bands as discovered by Dunlavy [17]. When fed by a second untuned loop, this antenna will exhibit a nearly constant feed point impedance over a three or four to one bandwidth by simply adjusting the capacitor to the desired resonant frequency. The reactive part of the loop impedance is inductive, where the inductance is given by $\{Z_L\} = \omega L$, so ignoring the higher-order terms

$$L = \frac{\eta_0 kb\left[\ln\left[\frac{8b}{a}\right] - 2\right]}{\omega} \tag{45}$$

which with the substitution $\eta_0 k/\omega = \mu_0$ becomes

$$L = \mu_0 b\left[\ln\left[\frac{8b}{a}\right] - 2\right] \tag{46}$$

The capacitance required to resonate this small loop at frequency f is

$$C = \frac{1}{(2\pi f)^2 L} \tag{47}$$

The loop may be coupled to a radio circuit in many different ways, including methods given in Refs. 17 and 18. When used in transmitter applications, the small-loop antenna is capable of impressing a substantial voltage across the resonating capacitor. For a power P delivered to a small loop with unloaded Q of Eq. (23) and with resonating the reactance X_C given by the reactive part of Eq. (22), it is easy to show that the peak voltage across the resonating capacitor is

$$V_p = \sqrt{X_C Q P} \tag{48}$$

by recognizing that

$$V_p = \sqrt{2} I_{\text{RMS}} X_C \tag{49}$$

where I_{RMS} is the total RMS loop current

$$I_{\text{RMS}} = \sqrt{\frac{P}{\text{Re}\{Z_{\text{loop}}\}}} \tag{50}$$

along with Q at the resonant frequency in Eq. (23).

Transmitter power levels as low as one watt delivered to a moderately efficient small-diameter ($\lambda/100$) loop can result in peak values of several hundred volts across the resonating capacitor. This is not intuitively expected; the small loop is often viewed as a high current circuit that is often described as a short-circuited ring. However, because it is usually implemented as a *resonant circuit* with a resonating capacitor, it can also be an extremely high-voltage circuit as will be shown below. Care must be exercised in selecting the voltage rating of the resonating capacitor even for modest transmitting power levels, just as care must be taken to keep resistive losses low in the loop structure.

As an example, consider the Q and bandwidth of a small-loop antenna: $2b = 10$ cm in diameter, resonated by a series capacitor and operating at 30 MHz. The example loop is constructed of $2a = 1$ cm diameter copper tubing with conductivity $\sigma = 5.7 \times 10^7$ S/m. The resistance per unit length of round wire of diameter $2a$ with conductivity σ is

$$R_s = \frac{1}{2\pi a \delta_s \sigma} = \frac{1}{2\pi a}\sqrt{\frac{\omega\mu_0}{2\sigma}} \tag{51}$$

where δ_s is the skin depth for good conductors and ω is the radian frequency and $\mu_0 = 4\pi \times 10^{-7}$ H/m is the permeability of free space, so $R_s = 0.046\ \Omega$. From Eq. (22) the loop impedance is $Z = 0.00792 + j71.41$. Hence the loop efficiency can be found by comparing the loop radiation resistance with loss resistance. The loop efficiency is $R_s/(R_s + \mathrm{Re}\{Z\}) = 0.147$ or 14.7%. From Eqs. (46) and (47) we find the resonating capacitance $C = 74.3\ \mu$F. From Eqs. (48)–(50) we see that if one watt is supplied to the loop, the peak voltage across the resonating capacitor is 308 V, and that the loop current 4.3 A. The *resonated* loop is by no means the "low impedance" structure that we normally imagine it to be.

3.4. The Rectangular Loop

Pager and other miniature receiver antennas used in the 30–940 MHz frequency range are most often implemented as electrically small rectangular loops. For a rectangle dimensioned $b_1 \times b_2$ of comparable length, and constructed with $2a$-diameter round wire, the loop impedance is given in Ref. 19 as

$$Z_{\mathrm{rect}} = \frac{\eta_0}{6\pi}(k^2A)^2 + j\frac{\eta_0}{\pi}\left[b_1 \ln\left[\frac{2A}{a(b_1+b_c)}\right] + \left[b_2 \ln\left[\frac{2A}{a(b_2+b_c)}\right] + 2(a + b_c - b_1 - b_2)\right]\right] \tag{52}$$

where $A = b_1 b_2$ and $b_c = (b_1^2 + b_2^2)^{1/2}$. The loss resistance is found by multiplying R_s in Eq. (51) by perimeter length of the loop, $2(b_1 + b_2)$. For a given antenna size the lowest loss occurs for the circular loop.

3.5. The Quad Loop Antenna

The quad loop antenna, sometimes called the *cubical quad*, was developed by Clarence C. Moore in the 1940s as a replacement for a four element parasitic dipole array (Yagi–Uda array). The dipole array exhibited corona arcing at the element tips severe enough to damage the antenna when operated at high power levels (10 kW) in a high-altitude (10,000-ft) shortwave broadcasting application in the 25-m band. Moore sought an antenna design with "no tips" that would support extremely high electric field strengths, which caused the destructive arcing. His solution was a one-wavelength perimeter square loop, and later with a loop director element as shown in Fig. 11. The configuration exhibited no arcing tendencies, and a new short wave antenna configuration was born.

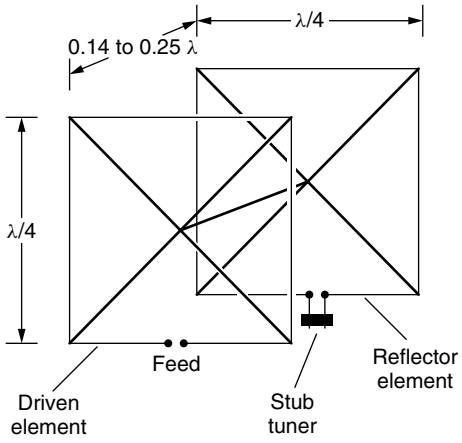


Figure 11. Two-element loop array.

As shown in Fig. 11, the driven element is approximately one quarter-wavelength on an edge. Actually, resonance occurs when the antenna perimeter is about 3% greater than a wavelength. The reflector element perimeter is approximately 6% larger than a wavelength, and may be implemented with a stub tuning arrangement. Typical element spacing is between 0.14 λ and 0.25 λ. The directivity of a quad loop is approximately 2 dB greater than that of a Yagi antenna with the same element spacing.

BIOGRAPHY

Kazimierz "Kai" Siwiak received his B.S.E.E. and M.S.E.E. degrees from the Polytechnic Institute of Brooklyn and his Ph.D. from Florida Atlantic University, Boca Raton, Florida. He designed radomes and phased array antennas at Raytheon before joining Motorola, where he received the Dan Noble Fellow Award for his research in antennas, propagation, and advanced communications systems. In 2000, he joined Time Domain Corporation to lead strategic technology development. He has lectured and published internationally; and holds more than 70 patents worldwide, including 31 issued in the United States. He was awarded Paper of the Year by IEEE–VTS and has authored, *Radiowave Propagation and Antennas for Personal Communications*, (Artech House), now in second edition, and contributed chapters to several other books and encyclopedias.

BIBLIOGRAPHY

1. *IEEE Standard Definitions of Terms for Antennas*, IEEE Std 145-1993, SH16279, March 18, 1993.
2. K. Siwiak, *Radiowave Propagation and Antennas for Personal Communications*, 2nd ed., Artech House, Norwood, MA, 1998.
3. G. J. Burke and A. J. Poggio, *Numerical Electromagnetics Code (NEC)—Method of Moments*, Lawrence Livermore Laboratory, NOSC Technical Document 116 (TD 116), Vols. 1 and 2, Jan. 1981.

4. H. C. Pocklington, Electrical oscillations in wires, *Proc. Cambridge Physical Society*, London, 1897, Vol. 9, pp. 324–333.
5. E. Hallén, Theoretical investigation into transmitting and receiving qualities of antennae, *Nova Acta Regiae Soc. Ser. Upps.* **II**(4): 1–44 (1938).
6. J. E. Storer, Impedance of thin-wire loop antennas, *Trans. AIEE* **75**(4): 609–619 (1965).
7. T. T. Wu, Theory of the thin circular antenna, *J. Math. Phys.* **3**: 1301–1304 (Nov.–Dec. 1962).
8. Q. Balzano and K. Siwiak, The near field of annular antennas, *IEEE Trans. Vehic. Technol.* **VT36**(4): 173–183 (Nov. 1987).
9. Q. Balzano and K. Siwiak, Radiation of annular antennas, *Correlations* (Motorola Eng. Bull., Motorola Inc., Schaumburg, IL, USA) **VI**(2): (1987).
10. C. A. Balanis, *Advanced Engineering Electromagnetics*, Wiley, New York, 1989.
11. R. E. Collin, *Antennas and Radiowave Propagation*, McGraw-Hill, New York, 1985.
12. E. C. Jordan and K. G. Balmain, *Electromagnetic Waves and Radiating Systems*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1968.
13. R. Cohen and B. N. Taylor, The 1986 CODATA recommended values of the fundamental physical constants, *J. Res. Natl. Bureau Stand.* **92**(2): (March–April 1987).
14. U.S. Patent 1,577,421 (March 16, 1926), L. A. Hazeltine, Means for eliminating magnetic coupling between coils.
15. R. W. P. King and C. W. Harrison, Jr., *Antennas and Waves: A Modern Approach*, MIT Press, Cambridge, MA, 1969.
16. R. Courant and D. Hibert, *Methods of Mathematical Physics*, Interscience, New York, 1953.
17. U.S. Patent 3,588,905 (June 28, 1971), J. H. Dunlavy, Jr., Wide range tunable transmitting loop.
18. T. Hart, Small, high-efficiency loop antennas, *QST J. ARRL* 33–36 (June 1986).
19. K. Fujimoto, A. Henderson, K. Hirasawa, and J. R. James, *Small Antennas*, Wiley, New York, 1987.

LOW-BIT-RATE SPEECH CODING

Miguel Arjona Ramírez
Mario Minami
University of São Paulo
São Paulo, Brazil

1. INTRODUCTION

Speech coders were first used for encrypting the speech signal as they still are today for secure voice communications. But their most important use is bit rate saving to accommodate more users in a communications channel such as a mobile telephone cell or a packet network link. Alternatively, a high-resolution coder or a more elaborate coding method may be required to provide for a higher-fidelity playback.

Actually, the availability of ever broader-band connection and larger-capacity media has led some to consider speech coding as unnecessary but the increasing population of transmitters and the increasingly rich content have taken up the "bandwidth" made available by the introduction of broadband services.

Further, coding may be required to counter the noise present in the communication channel, such as a wireless connection, or the decay of the storage media, such as a magnetic or optical disk. In fact, such a coding, called *channel coding*, will increase the total bit rate, and this is usually on a par with encryption. In contrast, the coding mentioned before is called *source coding* and will be dealt with almost exclusively below.

The speech signal is an analog continuous waveform, and any digital representation of it incurs a distortion or lack of fidelity, which is irrelevant for high-fidelity rendering. High-fidelity representations are obtained by filtering the signal within a sufficiently wide frequency band, sampling it at regular intervals and then quantizing each amplitude so obtained with a large number of bits. This kind of direct digital coding is called *pulse-code modulation* (PCM). The sampling operation is reversible if properly done, and the large number of bits for quantizer codes makes it possible to have a large number of closely spaced coding levels, reducing quantization distortion.

Since human hearing has a finite sensitivity, a sufficiently fine digital representation may be considered "transparent" or essentially identical to the original signal. In the case of a general audio signal, a bit rate of 706 kbit/s per channel, compact-disk (CD) quality, is usually considered transparent, while for telephone speech 64 kbps (kilobits per second) is taken as toll quality (Table 1). Even though it is rather elusive to impose a range for low-bit-rate speech coding as it is a moving target, it seems that nowadays it is best bounded by 4 kbps from above, given the longlasting effort to settle for a toll quality speech coder at that rate at the ITU-T [1,2], and it is bounded by $\approx$1 kbps from below by considering mainly the expected range of leading coding techniques at the lower low-rate region and the upper very-low-rate region [3]. A very good and comprehensive reference to speech coding [4] located low rate between 2.4 kbps and 8 kbps just some years ago.

Table 1. Bit Rates of Typical Acoustic Signals

	Bandwidth (Hz–kHz)	Sampling Frequency	Bits per Sample	Bit Rate (kbps)
Narrowband speech	300–3.4	8.0	8	64
Wideband speech	50–7.0	16.0	14	224
Wideband audio (DAT format)	10–20.0	48.0	16	768
Wideband audio (CD format)	10–20.0	44.1	16	706

2. SPEECH MODELING FOR LOW-RATE SPEECH CODING

Speech is a time-varying signal that may be considered stationary during segments of some tens of milliseconds in general. For these segments, usually called *frames*, an overall characterization is often made by using a spectral model. Complementarily, the energy is imparted to a synthesis filter, which embodies the estimated spectral model, by an excitation signal also carrying more details of the fine structure of the signal spectrum, or else the spectral model may be sampled at selected frequencies or integrated over selected frequency bands in order to define a proper reconstructed signal. In addition, the incorporation into the excitation model of the requisite interpolation for the process of synthesis further extends it into the time–frequency domain.

2.1. Predictive Coders

During the first half of the twentieth century, filterbanks were used for synthesizing speech since the first voice coder or "vocoder" developed by Dudley. The major difficulty in vocoding was the separation of vocal source behavior from vocal-tract behavior in order to drive a source–filter model for synthesis. A didactic taxonomy of parametric coders is given by Deller et al. [5].

A manageable and accurate acoustical model of speech production was proposed by Fant in 1960, and a good approximation to it is provided by the linear prediction (LP) model. The LP model for speech analysis was originally proposed by Itakura and Saito in 1968 and Atal and Hanauer in 1971 [6], whose spectral models are short-term stationary and nonstationary, respectively. The stationary LP spectral model is the frequency response of

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{1}$$

whose magnitude may be interpreted as a fit to the envelope of the short-term log spectrum of the signal as shown in Fig. 1. The order p, of the LP model has to be high enough to enable it to adjust to the overall shape of the spectrum, and the gain factor G allows an energy matching between the frequency response of the model and the spectrum of the signal. The LP model is particularly biased toward the peaks of the signal spectrum as opposed to the valleys and is particularly useful as a smooth peak-picking template for estimating the formants, sometimes not at likely places at first glance, like the second formant frequency in Fig. 1.

The excitation model proposed by Itakura and Saito combines two signal sources as shown in Fig. 2 whose relative intensities may be controlled by the two attenuation factors $U^{1/2}$ and $V^{1/2}$, which are interlocked by the relation

$$U + V = 1 \tag{2}$$

The pulse source, obtained for $V = 1$ and $U = 0$, is useful for generating voiced speech. In this mode, besides the gain factor G, the pulse repetition rate P has to be

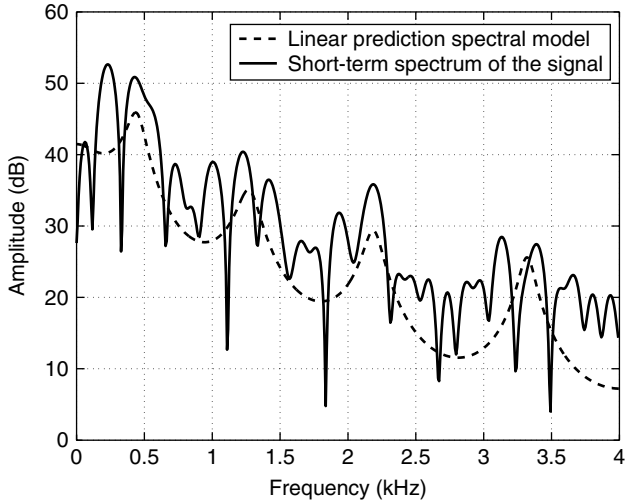


Figure 1. Linear prediction spectral fit to the envelope of the short-term log spectrum of the signal.

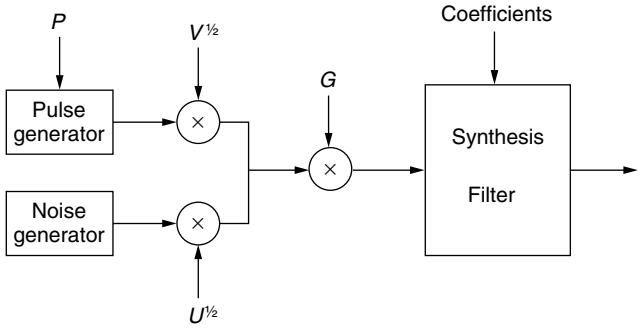


Figure 2. Mixed source–filter model for speech synthesis.

controlled. It is obtained in the coder as the pitch period of the speech signal through a pitch detection algorithm. The detected pitch period value may not be appropriate in many situations that may occur because of the quasiperiodic nature of voiced speech, the interaction of fundamental frequency (F_0) with the first formant or missing lower harmonics of F_0. On the other hand, for unvoiced speech the gain factor G is sufficient to match the power level of the pseudorandom source along with $U = 1$ and $V = 0$.

A better mixed excitation is produced by the mixed-excitation linear prediction (MELP) coder, which, besides combining pulse and noise excitations, is able to yield periodic and aperiodic pulses by position jitter [7]. Further, the composite mixed excitation undergoes adaptive spectral enhancement prior to going through the synthesis filter to produce the synthetic signal that is applied to the pulse dispersion filter.

2.2. Sinusoidal Coders

The voiced mode of speech production motivates the sine-wave representation of voiced speech segments by

$$s(n) = \sum_{k=1}^{K} A_{\cos}(\omega_k n + \phi_k) \tag{3}$$

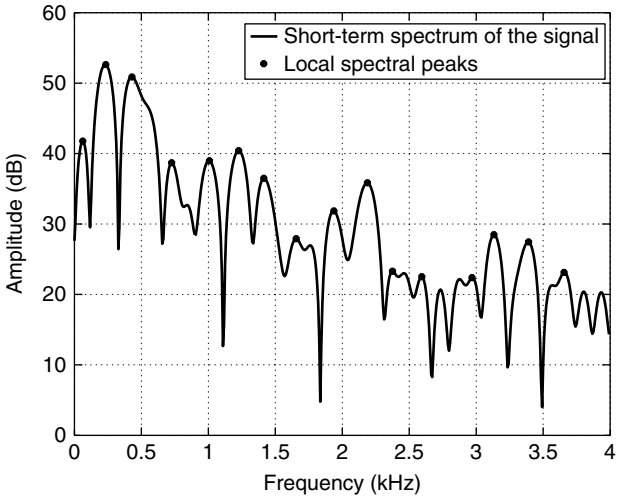


Figure 3. Short-term log spectrum of the signal with selected local peaks.

where A_k and ϕ_k are respectively the amplitude and phase of oscillator k, associated with the ω_k frequency track. This model quite makes sense in view of the spectrum of a voiced segment as can be seen in Fig. 3. As suggested in this figure, the peak frequencies $\{\omega_k,\ k = 1, 2, \ldots, K\}$ may be extracted and used as the oscillator frequencies in Eq. (3). For a strict periodic excitation model, $\omega_k = k\omega_0$, that is, the peak frequencies are equally interspaced and we have the so-called harmonic oscillator model. However, not all sinusoidal coders subscribe to this model because, by distinguishing small deviations from harmony, tonal artifacts may be guarded against. But the harmonic model is more amenable to low-rate implementation; thus other techniques have to be used to forestall the development of "buzzy" effects, which arise as a consequence of the forced additional periodicity.

The amplitudes may be constrained to lie on an envelope fit to the whole set of amplitudes, thereby enabling an efficient vector quantization of the amplitude spectrum. This amplitude model is compatible with the linear prediction filter described in Section 2.1, and the efficient quantization methods available for it may be borrowed, as is done for the sinusoidal transform coder (STC) [8].

Equation (3) may also be used for synthesizing unvoiced speech as long as the phases are random. In order to reduce the accuracy required of the voicing decision, a uniformly distributed random component is added to the phase of the oscillators with frequency above a voicing-dependent cutoff frequency in the STC as the lower harmonics of F_0 are responsible for the perception of pitch. In the multiband excitation (MBE) coder, the band around each frequency track is defined as either voiced or unvoiced, and Eq. (3) is not used for unvoiced synthesis; instead, filtered white noise is used. The bands are actually obtained after the signal has been windowed, and, as the windows have a finite bandwidth, this brings about a similarity of the sinusoidal coder with subband coders.

For low-rate coding, there is not enough rate for coding the phases, and phase models have to be used by the synthesizer such as the zero-phase model and the minimum-phase model. When there is a minimum-phase spectral model as in the latter case, the complex amplitude is obtained at no additional cost by sampling its frequency response as

$$H(e^{j\omega_k}) = A_k^{(r)} e^{j\phi_k^{(r)}} \tag{4}$$

where $A_k^{(r)}$ and $\phi_k^{(r)}$ are the reconstructed amplitude and phase of frequency track ω_k, respectively.

2.3. Waveform-Interpolation Coders

Waveform-interpolation coders usually apply linear prediction for estimating a filter whose excitation is made by interpolation of characteristic waveforms. Characteristic waveforms (CWs) are supposed to represent one cycle of excitation for voiced speech. The basic idea for the characteristic waveform stems from the Fourier series representation of a periodic signal, whose overtones are properly obtained by a Fourier series expansion. Therefore, the CW encapsulates the whole excitation spectrum, provided the signal is periodic. The rate of extraction of CWs may be as low as 40 Hz for voiced segments, as these waveforms are slowly varying in this case. On the other hand, for unvoiced segments the rate of extraction may have to be as high as 500 Hz but each segment may be represented with lower resolution [9].

The length of sampled characteristic waveforms varies as the pitch period. Therefore, their periods have to be normalized and aligned before coding for proper phase tracking. A continuous-time notation encapsulates a length normalization and the time-domain CW extraction process so that a two-dimensional surface may be built. The normalization of CW length is achieved by stretching or shrinking the waveforms to fit them within a normalized period of 2π radians. This normalized time within a period is referred to as the phase (ϕ). Assuming that linear prediction analysis has been performed and that the prediction residual has been determined for CW extraction and Fourier series representation, above and below the time–phase plane undulates the characteristic surface

$$u(t, \phi) = \sum_{k=1}^{K} \alpha_k(t) \cos(k\phi) + \beta_k(t) \sin(k\phi) \tag{5}$$

For the sake of coding efficiency, it is convenient to decompose the characteristic surface into a slowly evolving waveform (SEW) and a rapidly evolving waveform (REW). The SEW may be obtained by lowpass filtering $u(t, \phi)$ along the t axis as shown in Fig. 4 and represents the quasiperiodic component of speech excitation, whereas the REW may be obtained by highpass filtering $u(t, \phi)$ along the t axis, representing the random component of speech excitation. Both components must add up to the original surface:

$$u(t, \phi) = u_{\mathrm{SEW}}(t, \phi) + u_{\mathrm{REW}}(t, \phi) \tag{6}$$

Characteristic waveforms may be represented by means other than a Fourier series but in the latter case they may be compared to sinusoidal coders, having smaller

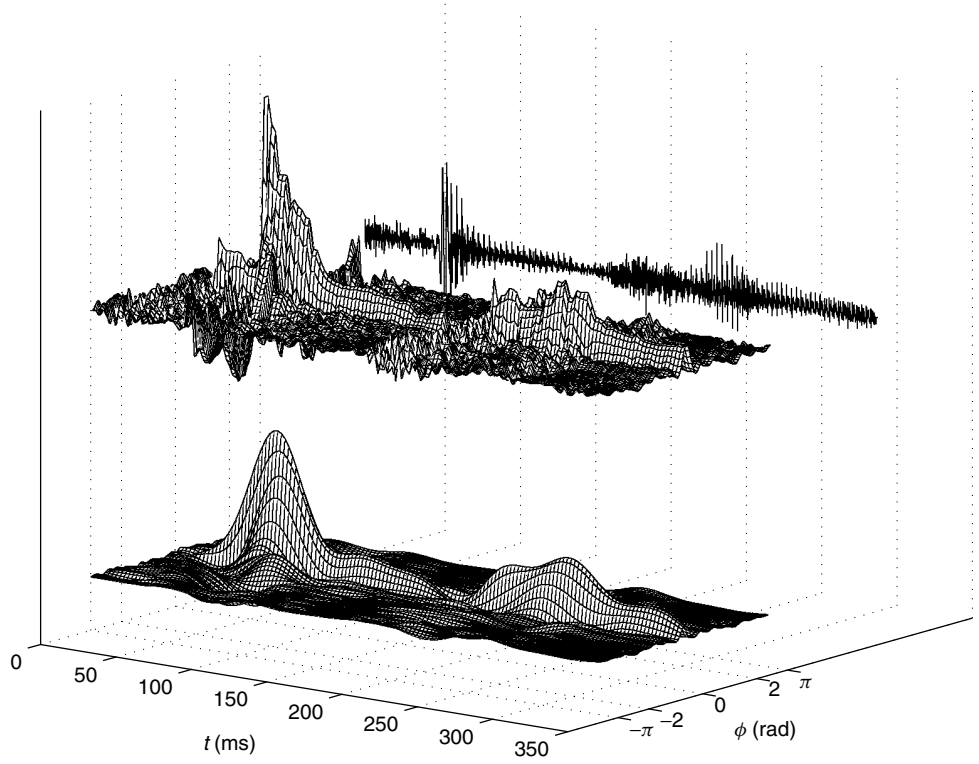


Figure 4. Characteristic surface for WI coding the residual signal given behind whose underlying CWs have been extracted at a 400 Hz rate. Its SEW component is also shown below, which has been obtained by lowpass filtering the characteristic surface along the time axis with a cutoff frequency of 20 Hz.

interpolation rates due to a more flexible time–frequency representation and to a higher resolution in time. For a common framework that encompasses both sinusoidal coding and waveform interpolation, please refer to Ref. 10, where the issue of perfect reconstruction in the absence of quantization errors is brought to bear.

3. PARAMETER ESTIMATION FROM SPEECH SEGMENTS

The linear prediction model was introduced in the last section along with the simplest excitation types for time-domain implementation, the frequency-domain parametric models of greater use for low-bit-rate coders and a harmonic excitation model, including waveform interpolation. In this section a more detailed description is provided of the structures used to constrain the excitation and the algorithms used for estimating its parameters. The segmentation of the speech signal for its analysis is complemented by its concatenation in the synthesis phase.

Although the initial goal was a medium bit rate range from 8 to 16 kbps, a different approach has come to be used for coding the excitation, called *code-excited linear prediction* (CELP) [11]. The two most important concepts in CELP coding are (1) an excitation quantization by sets of consecutive samples, which is a kind of vector quantization (VQ) of the excitation, and (2) a search criterion based on the reconstruction error instead of the prediction error or differential signal. Figure 5 has been drawn stressing these main distinguishing features.

A CELP coder is provided with a finite set of codevectors to be used for reconstructing each segment or subframe of the original signal. A collection of M codevectors is said to be a codebook of size M. Prior to searching the excitation, a filter is estimated through LP analysis (see Section 2.1) to have a frequency response matching the short-term spectral envelope of a block of the original signal called a "frame." Each frame typically consists of two to four excitation subframes, and the synthesis filter is determined for each subframe by interpolation from the LP filters of neighboring frames. As shown in Fig. 5, each codevector $\mathbf{c}_k$ in turn, for $k = 1, 2, \ldots, M$ is filtered by the synthesis filter

$$H(z) = \frac{1}{1 - P(z)} \tag{7}$$

generating all around the encoding loop a reconstruction error vector ε_k. This process of determining the signal to be synthesized within the coder is called the *analysis-by-synthesis method*. It allows the coder to anticipate the best strategy constrained to the situation that the synthesizer will face. Thus, the minimum square reconstruction error is identified as

$$i = \underset{k=1,2,\ldots,M}{\operatorname{argmin}} \{\|\varepsilon_k\|^2\} \tag{8}$$

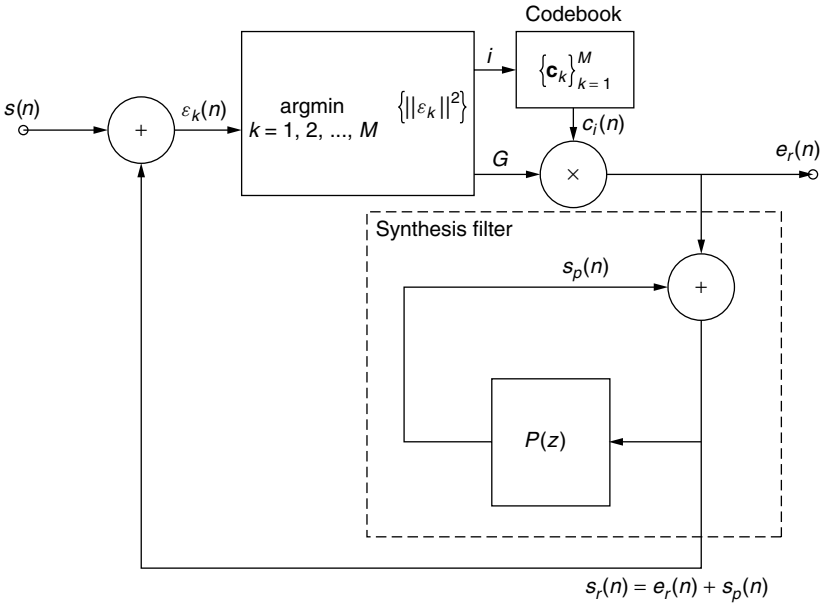


Figure 5. Conceptual block diagram for CELP coding.

after an exhaustive search all through the codebook and the actual excitation is delivered as the scaled version

$$\mathbf{e}_r = G\mathbf{c}_i \tag{9}$$

of codevector $\mathbf{c}_i$, where the scale factor $G = G_i$ has been calculated to minimize the square reconstruction error $\|\varepsilon_i\|^2$ for codevector $\mathbf{c}_i$.

Actually, a CELP coder applies a perceptual spectral weighting to the reconstruction error prior to the minimization by means of the weighting filter, defined by a function of the adaptive synthesis filter as

$$W(z) = \frac{H(z/\gamma_2)}{H(z/\gamma_1)} \tag{10}$$

where $0 < \gamma_2 < \gamma_1 \leq 1$ are bandwidth expansion factors. A very usual combination of values is $\gamma_2 = 0.8$ and $\gamma_1 = 1$. Overall, the weighting filter serves the dual purpose of deemphasizing the power spectral density of the reconstruction error around the formant frequencies where the power spectrum of the signal is higher and emphasizing the spectral density of the error in between the formant frequencies where hearing perception is more sensitive to an extraneous error. Both actions come about as consequences of the frequency response of $W(z)$ in Fig. 6. In much the same way, in order to achieve a reconstructed signal with a higher perceptual quality, an open-loop postfilter is usually applied to the reconstructed signal, which is defined as a function of the synthesis filter as well (see Fig. 7).

Additionally, toll quality reconstruction can be achieved only if there is a rather precise means of imposing the periodicity of voiced speech segments on the reconstructed signal. This goal can be achieved by using a second adaptive codebook in the CELP coder. This adaptive codebook is fed on a subframe basis the composite coded excitation

$$e(n) = G_a c_a(n) + G_f c_f(n) \tag{11}$$

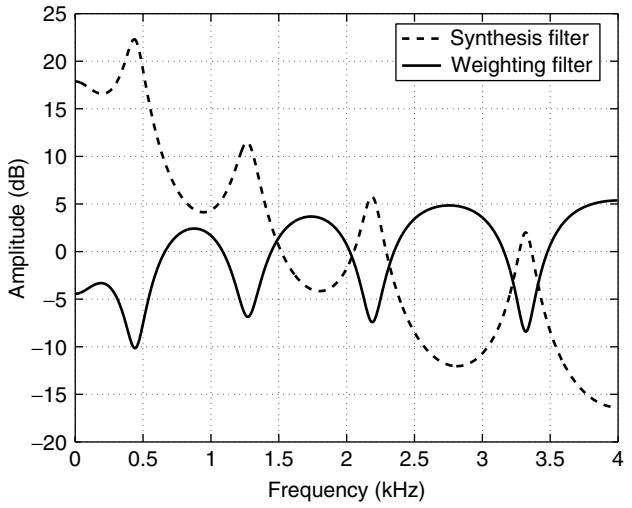


Figure 6. Frequency responses of synthesis filter and corresponding perceptual weighting filter.

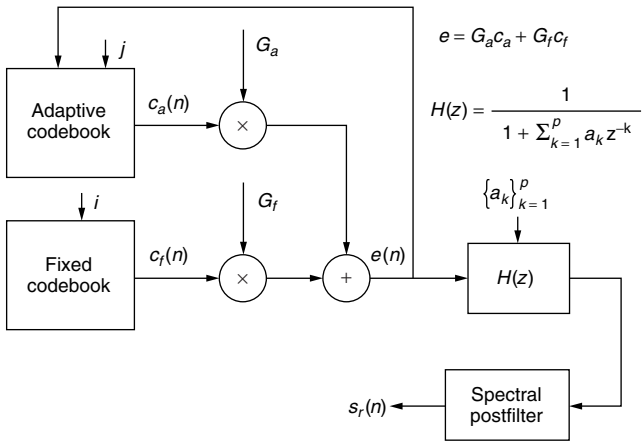


Figure 7. Two-codebook CELP synthesis model.

where $c_a(n)$ stands for the adaptive codevector with its gain factor G_a and $c_f(n)$ with its gain factor G_f represents the fixed excitation, depicted by the only codebook in Fig. 5. The enhanced synthesis model for this CELP coder is illustrated in Fig. 7.

Nonetheless, the fixed codebook structure and its search algorithms have been the target for developments leading to the widespread applicability of CELP coders. The fixed codebook in the original CELP coder was stochastically populated from samples of independent and identically Gaussian distributed vectors [11]. As the complexity of exhaustive searches through the codebook was overwhelming for the then-current signal processors, more efficient search methods were derived (discussed in Section 4), which required more structured codebooks such as the center-clipped and overlapped stochastic codebooks. Their searches have lower operational complexity due to the sparse amplitude distribution and the overlapped nature of their codevectors. The latter allows for the use of efficient search techniques originally developed for the adaptive codebook. Even more surprising, they enhance the speech quality as well [12] to a level considered good enough for secure voice and cellular applications at low to medium rates.

Meanwhile, predictive waveform coders borrow the idea of impulse excitation from parametric LP coders (see Section 2.1) in order to decrease the bit rate but with a twist to deliver higher quality, which involves the increase in the number of pulses per pitch period. A subframe of multipulse excitation is given by

$$e(n) = G \sum_{k=0}^{M-1} \alpha_k \delta(n - m_k), \quad n = 0, 1, \ldots, L-1 \qquad (12)$$

where M is the number of pulses per excitation subframe, L is the length of the subframe, α_k and m_k respectively represent individual pulse amplitude and position, and G is a common excitation vector gain. This new approach was called "multipulse excitation" and is very complex in its most general formulation [13]. Moreover, a constrained version of it, known by "regular pulse excitation with long-term predictor" (RPE-LTP), was adopted for the Global System for Mobile Communications (GSM) full-rate standard coder for digital telephony and is notable for its low complexity [14].

This kind of excitation was further structured and inserted into a CELP coder. Pulse positions were constrained to lie in different tracks, which cover in principle all the positions in the excitation subframe, whereas pulse amplitudes α_k were restricted to either plus or minus one. The latter feature and its conceptual connection to error-correction codes has established the name "algebraic CELP" for this kind of excitation. These deterministic sparse codebooks made their entrance into standard speech coding with the G.729 conjugate structure, algebraic CELP (CS-ACELP) coder [15]. A general ACELP position grid is given in Table 2 for an M-pulse codebook over an L-sample subframe.

As the bit rate is decreased, further modeling and classification of the signal has to be done at the encoder in order to keep speech quality about the same. For instance, the pitch synchronous innovation CELP (PSI-CELP) coder adapts the fixed random codevectors in voiced frames to have periodicity [16].

Table 2. ACELP Position Grid for *M*-Pulse Tracks over an *L-Sample* Subframe

Track	Positions				
0	0	M	$2M$	$\cdots$	$L-M$
1	1	$M+1$	$2M+1$	$\cdots$	$L-M+1$
2	2	$M+1$	$2M+2$	$\cdots$	$L-M+2$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$M-1$	$M-1$	$2M-1$	$3M-1$	$\cdots$	$L-1$

Surprisingly, the analysis-by-synthesis operation of CELP is proving capable of delivering toll-quality speech at lower rates when generalized to allow for a mixture of open-loop and closed-loop procedures [2] where parameters and excitation are determined in an open-loop fashion for clearly recognizable subframe types such as stationary periodic or voiced segments and closed-loop algorithms are used for unvoiced or transient segments. Because of the scarcity of bits for representing the excitation, it makes sense to predistort the target vector for closed-loop searches when it is clearly voiced since it becomes easier to match a codevector to it. The predistortion has to be perceptually transparent such as the time warping described in Ref. 17.

In a different trend, the development of text-to-speech (TTS) systems has been moving away from the rule-based, expert system approach to the new framework of concatenative synthesis, based on model fitting with statistical signal processing [18]. In rule-based systems subword speech units are designed as well as rules for concatenating them that take into account the coarticulation between neighboring units as well as their exchange for allophonic variations. On the other hand, concatenative synthesis systems are based on the acquisition of a large database of connected speech from an individual speaker containing instances of coarticulation between all possible units. For the latter systems, the synthesis consists of selecting the largest possible string of original database subunits, thereby borrowing their natural concatenation. The final postprocessing stage of the TTS adjusts the prosody of the synthetic signal, mostly by pitch and timescale modifications. For segment selection, a concatenative synthesizer uses both an acoustic cost within each segment and a concatenation cost between consecutive segments [3]. If the input feature vector sequence $\mathbf{F} = \mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N$ is to be synthesized by the unit sequence $\mathbf{U} = \mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N$, the acoustic cost may be defined by

$$J_A(\mathbf{f}_m, \mathbf{u}_m) = \sum_{k=1}^{K} (f_{m,k} - u_{m,k})^2 \qquad (13)$$

for segment m, where k indices through the K features are selected for comparison, normally the spectral representation of the subunits, and the concatenation cost

may be calculated by

$$J_C(\mathbf{u}_{m-1}, \mathbf{u}_m) = \sum_{k=1}^{K} (u_{m-1,k} - u_{m,k})^2 \tag{14}$$

The best subunit sequence is selected by minimization of the total cost $J(\mathbf{F}, \mathbf{U})$ whose simplest definition is

$$J(\mathbf{F}, \mathbf{U}) = \sum_{m=1}^{N} J_A(\mathbf{f}_m, \mathbf{u}_m) + \sum_{m=2}^{N} J_C(\mathbf{u}_{m-1}, \mathbf{u}_m) \tag{15}$$

With the use of these kinds of cost measures in their analysis, concatenative synthesizers are becoming more similar to speech coders.

4. LOW-RATE CODING APPROACHES

Speech coding allows more users to share a communications channel such as a mobile telephone cell or a packet network link and is concerned with the economical representation of a speech signal with a given distortion for a specified implementation complexity level. Traditionally, a fixed bit rate and an acceptable maximum distortion are specified. More generally, the required maximum bit rate or the acceptable maximum distortion level may be specified. Actually, for modern cellular or packet communications, sometimes the bit rate may be dictated by channel traffic constraints, requiring variable-bit-rate coders.

Objective fidelity measures such as the segmental signal-to-noise ratio (SNRSEG) are very practical for coder development, while more perceptual methods such as objective distortion measures, including the perceptual speech quality measure (PSQM) [19], which use to advantage the limitations of the human ear, may be used instead. But subjective the opinion of human listeners is still the best gauge of fidelity and may be assessed by the *mean opinion score* (MOS), obtained in formal listening tests where each listener classifies the speech stimulus on the 5-point scale shown in Table 3.

Coder complexity constrains the possibilities of rate distortion tradeoff. Its major component is operational complexity, liable to be measured in million instructions per second (MIPS) [20]. An artistic conception of the fidelity versus rate behavior of low-rate coders for two levels of complexity is presented in Fig. 8, anchored by some real coder test points, listed in Table 4. It should be mentioned that these fidelity curves pass through a kind of "knee" around the 4 kbps rate, where they evolve at a lower slope, eventually reaching a virtual plateau at high rates [21].

Low-bit-rate implementations of models tested at higher rates need compensation for the loss of resolution or reduction of parameters, whereas very-low-bit-rate

Table 3. Quality Scale for Subjective Listening Rating

Quality	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

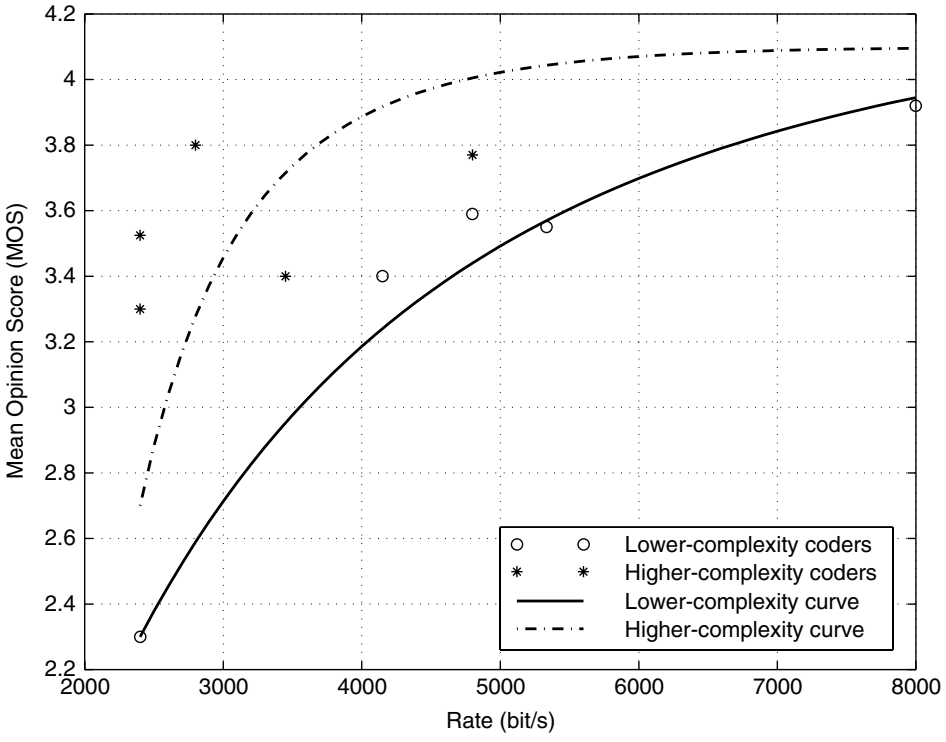


Figure 8. Conception of the fidelity versus rate behavior of low-rate speech coders for two levels of complexity, anchored by some real coder test points, listed in Table 4.

Table 4. Speech Quality and Operational Complexity of Some Selected Coders[a]

Coder	Bit rate (kbps)	Quality (MOS)	Complexity (Mips)[b]	Ref.
LPC-10e, FS-1015	2.40	2.30	8.7	37
MELP, FS-1017	2.40	3.30	20.4	37
EWI	2.80	~3.80	~30.0	33,35,38
PSI-CELP, RCR PDC half-rate	3.45	~3.40	23.0	14,16,38,39
IMBE, INMARSAT-M System	4.15	3.40	7.0	4,14
CELP, FS-1016	4.80	3.59	17.0	37,40
STC	4.80	3.53	~25.0	8
WI	4.80	3.77	~25.0	40
ACELP, G.723.1	5.33	3.55	16.0	33,41
CS-ACELP, G.729	8.00	3.92	20.0	38,41

[a] *Caution*: These performance and complexity figures were obtained under different test and implementation conditions and should be used only as a first guess in comparisons. Tilde (~) indicates estimate.
[b] Million instructions per second.

implementations admit refinements when upgraded to the low-rate range. In general, low-rate implementations require higher complexity algorithms and incur longer algorithmic delay. But a reduction in complexity may render the original algorithm useful for a number of applications. This is one reason why a number of efficient search algorithms have been proposed since the inception of the CELP coder such as that due to Hernández-Gómez et al. [22], who proposed a residual-based preselection of codevectors and the efficient transform-domain search algorithms elaborated by Trancoso and Atal [23]. Another preselection of codevectors was proposed [24] on the basis of the correlation between the backward-filtered target vector and segments of codevectors. The latter efficient search was called "focused search" and was adopted for the reference ITU-T 8-kbps CS-ACELP coder [15] with an open-loop signal-selected pulse amplitude approach. This coder is used for transmitting voice over packet networks among other applications.

In fact, the acceptance of this family of coders is so wide that most of the second-generation digital cellular coders use it, including the Telecommunications Industry Association (TIA) IS641 enhanced full-rate (EFR) coder [25] and the IS127 enhanced variable-rate coder (EVRC) [26] as well as the GSM EFR coder [27]. In addition, a general-purpose efficient search algorithm for ACELP fixed excitation codebook has been proposed, the joint position and amplitude search (JPAS) [28], which includes a closed-loop sequential pulse amplitude determination, and a more efficient search for the EVRC [29] has been advanced as well. Also, a generalization of "algebraic pulses" by "algebraic subvectors" is the basis for the algebraic vector quantized CELP (AVQ-CELP) search, which enhances the IS127 coder and uses open-loop subvector preselection in order to make it more efficient [30].

As the bit rate is decreased below 6 kbps, ACELP coder quality degrades because of the uniform pulse density in the pulse position grid [31] and the high level of sparsity in the resulting excitation waveform. In an effort to push down the bit rate for ACELP applications, pulse dispersion techniques have been proposed [32,33]. The former closed-loop technique is incorporated in a partially qualified candidate for the ITU-T 4-kbps coder [2]. Furthermore, parametric coders such as MELP also implement pulse dispersion but as an open-loop enhancement in the decoder as mentioned in Section 2.1. Along with pulse dispersion, the pulse position in the grid should be changed adaptively since it will not be able to cover all the positions [31,34].

Another technique that holds promise for lower-bit-rate coding is target vector predistortion. Time-warping predistortions have already been proposed as mentioned in Section 3 and even used in the IS127 EVRC.

The segments coded open loop may use enhanced vocoderlike techniques such as those used in the MELP or sinusoidal coders or, alternatively, WI techniques with a partial use of analysis-by-synthesis methods [35].

The judicious application of these enhancement techniques requires classification of the signal into voice or silence. In the former case, the speech signal is classified into voiced and unvoiced stationary segments at least. Even the identification of transients may be required as a next step. Branching out further, speech classification might get down to subunits such as triphones, diphones, and phones. In these cases the segmentation is event-driven, similar to the method used for very-low-rate coding [36]. Nevertheless, one should bear in mind that irregular segmentation requires timescale modification as a postprocessing stage, which may introduce annoying artifacts into the reconstructed signal. So sometimes it may be wise to maintain regular frame-based segmentation even at very low rates in order to ensure a certain uniform quality level [3].

In conclusion, the CELP framework with some relaxed waveform matching constraints, allowing for perceptual quality preserving signal predistortion and more segments of simple parametric coding, is very likely to be able to achieve toll quality at 4 kbps. It is anticipated as well that

coders based on codebooks of sequences of speech subunits with properly defined distortion measures will also play an important role in advancing the toll quality frontier into the low-bit-rate range.

BIOGRAPHIES

Miguel Arjona Ramírez received the E.E. degree from Instituto Tecnológico de Aeronáutica (ITA), Brazil, in 1980 and the M.S.E.E. and Dr.E.E. degrees from University of São Paulo, Brazil, in 1992 and 1997, respectively.

In 1981, while studying at Philips International Institute, he worked with coding algorithms for a formant speech synthesizer at Philips Electronic Components and Materials Laboratories, The Netherlands.

He joined Itautec Informática S.A., São Paulo, Brazil, as a Full Development Engineer in 1982, eventually becoming an Engineering Development Group Leader for Interactive Voice Response (IVR) Systems in 1988.

He is currently Assistant Professor at Escola Politécnica, University of São Paulo, where he has conducted research on predictive speech coders since 1991. Dr. Arjona Ramírez became Senior Member of the IEEE in 2000. His research interests include signal compression, speech coding and recognition, and audio coding with applications to circuit and packet telephony.

Mario Minami received the B.S. degree in physics in 1989 from Physics Institute of University of Sao Paulo, Sao Paulo, Brazil, and the M.S. and Ph.D. degrees in electrical engineering from the Politechnic School of University of Sao Paulo, Sao Paulo, Brazil in 1993 and 1998, respectively. He joined the OSUC — Obras Sociais, Universitarias e Culturais in 1989-91 as a Research Engineer and professor. At OSUC he worked on the design and development of didactic DSP projects. Since 1993, he has been a Research Scientist at LPS-EPUSP — Signal Processing Laboratory of Politechnic School of University of Sao Paulo. His areas of interest are speech and speaker recognition, speech coding, database for speech applications and channel equalization algorithms.

BIBLIOGRAPHY

1. S. Dimolitsas, C. Ravishankar, and G. Schröder, Current objectives in 4-kb/s wireline-quality speech coding standardization, *IEEE Signal Process. Lett.* **1**(11): 157–159 (Nov. 1994).
2. J. Thyssen et al., A candidate for the ITU-T 4 kbit/s speech coding standard, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Salt Lake City, 2001, Vol. 2, pp. 681–684.
3. K.-S. Lee and R. V. Cox, A very low bit rate speech coder based on a recognition/synthesis paradigm, *IEEE Trans. Speech Audio Process.* **9**(5): 482–491 (July 2001).
4. A. S. Spanias, Speech coding: A tutorial review, *Proc. IEEE* **82**(10): 1541–1582 (Oct. 1994).
5. J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan, 1993, Chap. 7, pp. 459–487.
6. J. D. Markel and A. H. Gray, *Linear Prediction of Speech*, Springer, Berlin, 1976.
7. A. McCree et al., A 2.4 kbit/s MELP coder candidate for the new U. S. Federal Standard, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Atlanta, 1996, Vol. 1, pp. 200–203.
8. R. J. McAulay and J. F. Quatieri, Sinusoidal coding, in W. B. Kleijn and K. K. Paliwal, eds., *Speech Coding and Synthesis*, Elsevier Science, Amsterdam, 1995, pp. 121–173.
9. W. B. Kleijn and K. K. Paliwal, An introduction to speech coding, in W. B. Kleijn and K. K. Paliwal, eds., *Speech Coding and Synthesis*, Elsevier Science, Amsterdam, 1995, pp. 1–47.
10. W. B. Kleijn, A frame interpretation of sinusoidal coding and waveform interpolation, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Istanbul, 2000, Vol. 3, pp. 1475–1478.
11. M. R. Schroeder and B. S. Atal, Code-excited linear prediction (CELP): High quality speech at very low bit rates, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Tampa, 1985, Vol. 2, pp. 437–440.
12. W. B. Kleijn, D. J. Krasinski, and R. H. Ketchum, Fast methods for the CELP speech coding algorithm, *IEEE Trans. Acoust. Speech, Signal Process.* **38**(8): 1330–1342 (Aug. 1990).
13. B. S. Atal and J. R. Remde, A new model of LPC excitation for producing natural-sounding speech at low bit rates, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Paris, 1982, Vol. 1, pp. 614–617.
14. R. V. Cox, Speech coding standards, in W. B. Kleijn and K. K. Paliwal, eds., *Speech Coding and Synthesis*, Elsevier Science, Amsterdam, 1995, pp. 49–78.
15. R. Salami et al., Design and description of CS-ACELP, a toll quality 8 kb/s speech coder, *IEEE Trans. Speech Audio Process.* **6**(2): 116–130 (March 1998).
16. K. Mano et al., Design of a pitch synchronous innovation CELP coder for mobile communications, *IEEE J. Select. Areas Commun.* **13**(1): 31–40 (Jan. 1995).
17. W. B. Kleijn, R. P. Ramachandran, and P. Kroon, Generalized analysis-by-synthesis coding and its application to pitch prediction, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, San Francisco, 1992, Vol. 1, pp. 23–26.
18. Y. Sagisaka and N. Iwahashi, Objective optimization in algorithms for text-to-speech synthesis, in W. B. Kleijn and K. K. Paliwal, eds., *Speech Coding and Synthesis*, Elsevier Science, Amsterdam, 1995, pp. 685–706.
19. *Objective Quality Measurement of Telephone-Band (300–3400 Hz) Speech Codecs*, ITU-T Recommendation P.861, Aug. 1996.
20. P. Kroon, Evaluation of speech coders, in W. B. Kleijn and K. K. Paliwal, eds., *Speech Coding and Synthesis*, Elsevier Science, Amsterdam, 1995, pp. 467–494.
21. N. S. Jayant and P. Noll, *Digital Coding of Waveforms*, Prentice-Hall, Englewood Cliffs, NJ, 1984.
22. L. A. Hernández-Gómez, F. J. Casajús-Quirós, A. R. Figueiras-Vidal, and R. García-Gómez, On the behaviour of reduced complexity code-excited linear prediction (CELP), in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Tokyo, 1986, Vol. 1, pp. 469–472.
23. I. M. Trancoso and B. S. Atal, Efficient procedures for finding the optimum innovation in stochastic coders, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Tokyo, 1986, Vol. 4, pp. 2375–2378.

24. C. Laflamme et al., 16 kbps wideband speech coding technique based on algebraic CELP, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Toronto, 1991, Vol. 1, pp. 13–16.

25. T. Honkanen, J. Vainio, K. Järvinen, and P. Haavisto, Enhanced full rate codec for IS-136 digital cellular system, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Munich, 1997, Vol. 2, pp. 731–734.

26. *Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems*, TIA/EIA/IS-127, July 1996.

27. K. Järvinen et al., GSM enhanced full rate speech codec, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Munich, 1997, Vol. 2, pp. 771–774.

28. M. A. Ramírez and M. Gerken, Joint position and amplitude search of algebraic multipulses, *IEEE Trans. Speech Audio Process.* **8**(5): 633–637 (Sept. 2000).

29. H. Park, Efficient codebook search method of EVRC speech codec, *IEEE Signal Process. Lett.* **7**(1): 1–2 (Jan. 2000).

30. F. Liu and R. Heidari, Improving EVRC half rate by the algebraic VQ-CELP, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Phoenix, 1999, Vol. 4, pp. 2299–2302.

31. V. Cuperman et al., A novel approach to excitation coding in low-bit-rate high-quality CELP coders, *Proc. IEEE Workshop on Speech Coding*, Delavan, Wisconsin, 2000, pp. 14–16.

32. K. Yasunaga, H. Ehara, K. Yoshida, and T. Morii, Dispersed-pulse codebook and its application to a 4 kb/s speech coder, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Istanbul, 2000, Vol. 3, pp. 1503–1506.

33. M. A. Ramírez, Sparsity compensation for speech coders, *Proc. IEEE GLOBECOM*, San Antonio, 2001, Vol. 4, pp. 2475–2478.

34. T. Amada, K. Miseki, and M. Akamine, CELP speech coding based on an adaptive pulse position codebook, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Phoenix, 1999, Vol. 1, pp. 13–16.

35. O. Gottesman and A. Gersho, Enhanced waveform interpolative coding at low bit-rate, *IEEE Trans. Speech Audio Process.* **9**(8): 786–798 (Nov. 2001).

36. C. S. Xydeas and T. M. Chapman, Segmental prototype interpolation coding, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Phoenix, 1999, Vol. 4, pp. 2311–2314.

37. M. A. Kohler, A comparison of the new 2.4 kbps MELP federal standard with other standard coders, in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Munich, 1997, Vol. 2, pp. 1587–1590.

38. M. E. Perkins, K. Evans, D. Pascal, and L. A. Thorpe, Characterizing the subjective performance of the ITU-T 8 kb/s speech coding algorithm — ITU-T G.729, *IEEE Commun. Mag.* **35**(9): 74–81 (Sept. 1997).

39. K. Mano, Design of a toll-quality 4-kbit/s speech coder based on phase-adaptive PSI-CELP, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Munich, 1997, Vol. 2, pp. 755–758.

40. W. B. Kleijn and J. Haagen, Waveform interpolation for coding and synthesis, in W. B. Kleijn and K. K. Paliwal, eds., *Speech Coding and Synthesis*, Elsevier Science, Amsterdam, 1995, pp. 175–207.

41. R. V. Cox and P. Kroon, Low bit-rate speech coders for multimedia communication, *IEEE Commun. Mag.* **34**(12): 34–41 (Dec. 1996).

LOW-DENSITY PARITY-CHECK CODES: DESIGN AND DECODING

SARAH J. JOHNSON*
STEVEN R. WELLER†
University of Newcastle
Callaghan, Australia

1. INTRODUCTION

The publication of Claude Shannon's 1948 paper, "A mathematical theory of communication" [1], marked the beginning of coding theory. In his paper, Shannon established that every communication channel has associated with it a number called the channel *capacity*. He proved that arbitrarily reliable communication is possible even through channels that corrupt the data sent over them, but only if information is transmitted at a rate less than the channel capacity.

In the simplest case, transmitted messages consist of strings of 0s and 1s, and errors introduced by the channel consist of bit inversions: $0 \rightarrow 1$ and $1 \rightarrow 0$. The essential idea of forward error control coding is to augment messages to produce codewords containing deliberately introduced redundancy, or *check* bits. With care, these check bits can be added in such a way that codewords are sufficiently distinct from one another so that the transmitted message can be correctly inferred at the receiver, even when some bits in the codeword are corrupted during transmission over the channel.

While Shannon's noisy channel coding theorem establishes the existence of capacity-approaching codes, it provides no explicit guidance as to how the codes should be chosen, nor how messages can be recovered from the noise-corrupted channel output. The challenges to communicating reliably at rates close to the Shannon limit are therefore twofold: (1) to design sets of suitably distinct codewords and (2) to devise methods for extracting estimates of transmitted messages from the output of a noise-contaminated channel, and to do so without excessive decoder complexity.

In this article, we consider code design and decoding for a family of error correction codes known as *low-density parity-check* (LDPC) *block codes*. In the simplest form of a parity-check code, a single parity-check equation provides for the detection, but not correction, of a single bit inversion in a received codeword. To permit correction of errors induced by channel noise, additional parity checks can be added at the expense of a decrease in the rate of transmission. Low-density parity-check codes are a special case of such codes. Here "low density" refers to the sparsity of the parity-check matrix characterizing the

* Work supported by a CSIRO Telecommunications & Industrial Physics postgraduate scholarship and the Centre for Integrated Dynamics and Control (CIDAC).
† Work supported in part by Bell Laboratories Australia, Lucent Technologies, as well as the Australian Research Council under Linkage Project Grant LP0211210, and the Centre for Integrated Dynamics and Control (CIDAC).

code. Each parity-check equation checks few message bits, and each message bit is involved in only a few parity-check equations. A delicate balance exists in the construction of appropriate parity-check matrices, since excessive sparsity leads to uselessly weak codes.

First presented by Gallager in his 1962 thesis [2,3], low-density parity-check codes are capable of performance extraordinarily close to the Shannon limit when appropriately decoded. Codes that approach the Shannon limit to within 0.04 of a decibel have been constructed. Figure 1 shows a comparison of the performance of LDPC codes with the performance of some well-known error correction codes. The key to extracting maximal benefit from LDPC codes is *soft-decision* decoding, which starts with a more subtle model for channel-induced errors than simple bit inversions. Rather than requiring that the receiver initially make *hard decisions* at the channel output, and so insisting that each received bit be assessed as either 0 or 1, whatever is the more likely, soft-decision decoders use knowledge of the channel noise statistics to feed probabilistic (or "soft") information on received bits into the decorder.

The final ingredient in implementing soft-decision decoders with acceptable decoder complexity are *iterative* schemes that handle the soft information in an efficient manner. Soft iterative decoders for LDPC codes make essential use of *graphs* to represent codes, passing probabilistic messages along the edges of the graph. The use of graphs for iterative decoding can be traced to Gallager, although for over 30 years barely a handful of researchers pursued the consequences of Gallager's work. This situation changed dramatically with the independent rediscovery of LDPC codes by several researchers in the mid-1990s, and graph-based representations of codes are now an integral feature in the development of both the theoretical understanding and implementation of iterative decoders.

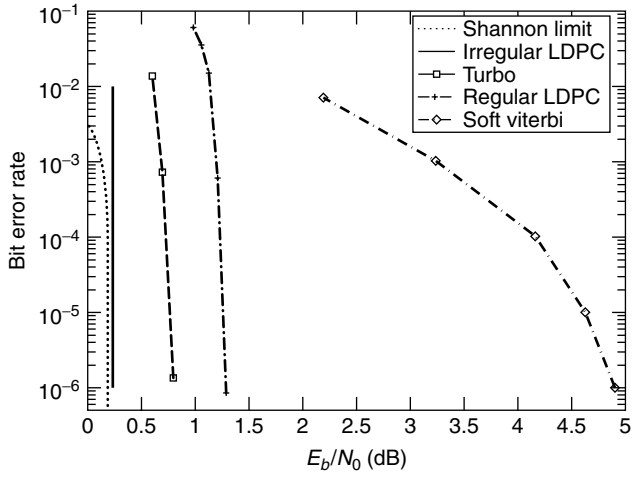


Figure 1. Bit error rate performance of rate-$\frac{1}{2}$ error correction codes on an additive white Gaussian noise channel. From right to left, soft Viterbi decoding of a constraint length 7 convolutional code; sum–product decoding of a regular Gallager code with blocklength 65,389 [8]; a Turbo code with $2 + 32$ states, 16,384-bit interleaver, and 18 iterations *http://www331.jp1.nasa.gov/public/TurboPerf.html*; sum–product decoding of a blocklength 10^7 optimized irregular code [16]; and the Shannon limit at rate $\frac{1}{2}$.

In this article, we begin by introducing parity checks and codes defined by their parity-check matrices. To introduce iterative decoding we present in Section 3 a hard-decision iterative algorithm that is not very powerful, but suggestive of how graph-based iterative decoding algorithms work. The soft-decision iterative decoding algorithm for LDPC codes known as *sum–product decoding* is presented in Section 4. Section 5 focuses on the relationship between the codes and the decoding algorithm, as expressed in the graphical representation of LDPC codes, and Section 6 considers the design of LDPC codes. The article concludes with a discussion of the connections of this work to other topics and future directions in the area.

2. LOW-DENSITY PARITY-CHECK CODES

2.1. Parity-Check Codes

The simplest possible error detection scheme is the single parity check, which involves the addition of a single extra bit to a binary message. Whether this parity bit should be a 0 or a 1 depends on whether even or odd parity is being used. In even parity, the additional bit added to each message ensures an even number of 1s in each transmitted codeword. For example, since the 7-bit ASCII code for the letter S is 1010011, a parity bit is added as the eighth bit. If even parity is being used, the value of the parity bit is 0 to form the codeword 10100110.

More formally, for the 7-bit ASCII plus even parity code, we define a codeword c to have the following structure:

$$c = c_1\, c_2\, c_3\, c_4\, c_5\, c_6\, c_7\, c_8$$

where each c_i is either 0 or 1, and every codeword satisfies the constraint

$$c_1 \oplus c_2 \oplus c_3 \oplus c_4 \oplus c_5 \oplus c_6 \oplus c_7 \oplus c_8 = 0 \qquad (1)$$

Here the symbol $\oplus$ represents modulo-2 addition, which is equal to 1 if the ordinary sum is odd and 0 if the ordinary sum is even. Whereas the inversion of a single bit due to channel noise can be easily detected with a single parity check [as (1) is no longer satisfied by the noise-corrupted codeword], this code is not sufficiently powerful to indicate which bit (or bits) was (were) inverted. Moreover, since any even number of bit inversions produces a word satisfying the constraint (1), any even numbers of errors go undetected by this simple code.

One measure of the ability of a code to detect errors is the *minimum distance* of the code. The *Hamming distance* between two codewords is defined as the number of bit positions in which they differ. For example, the codewords 10100110 and 10000111 differ in positions 3 and 8, so the Hamming distance between them is 2. The minimum distance of a code, $d_{\min}$, is defined as the smallest Hamming distance between any pair of codewords in the code. For the even parity code $d_{\min} = 2$, so the corruption of 2 bits in a codeword can result in another valid codeword and will consequently not be detected.

Detecting more than a single bit error calls for increased redundancy in the form of additional parity checks. To illustrate, suppose that we define a codeword c to have the following structure:

$$c = c_1\ c_2\ c_3\ c_4\ c_5\ c_6$$

where each c_i is either 0 or 1, and c is constrained by three parity-check equations:

$$\begin{array}{l} c_1 \oplus c_2 \oplus c_4 = 0 \\ c_2 \oplus c_3 \oplus c_5 = 0 \\ c_1 \oplus c_2 \oplus c_3 \oplus c_6 = 0 \end{array} \Leftrightarrow \underbrace{\begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}}_{H} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad (2)$$

In matrix form we have that $c = [c_1\ c_2\ c_3\ c_4\ c_5\ c_6]$ is a codeword if and only if it satisfies the constraint

$$Hc^T = 0 \quad (3)$$

where the *parity-check matrix*, H, contains the set of parity-check equations that define the code. To generate the codeword for a given message, the code constraints can be rewritten in the form

$$\begin{array}{l} c_4 = c_1 \oplus c_2 \\ c_5 = c_2 \oplus c_3 \\ c_6 = c_1 \oplus c_2 \oplus c_3 \end{array} \Leftrightarrow [c_1\ c_2\ c_3\ c_4\ c_5\ c_6] = [c_1\ c_2\ c_3] \underbrace{\begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}}_{G} \quad (4)$$

where bits c_1, c_2, and c_3 contain the 3-bit message, and parity-check bits c_4, c_5, and c_6 are calculated from the message. Thus, for example, the message 110 produces parity-check bits $c_4 = 1 \oplus 1 = 0$, $c_5 = 1 \oplus 0 = 1$, and $c_6 = 1 \oplus 1 \oplus 0 = 0$, and hence the codeword 110010. The matrix G is the *generator matrix* of the code. Substituting each of the $2^3 = 8$ distinct messages $c_1\ c_2\ c_3 = 000, 001, \ldots, 111$ into Eq. (4) yields the following set of codewords:

$$\begin{array}{cccc} 000000 & 001011 & 010111 & 011100 \\ 100101 & 101110 & 110010 & 111001 \end{array} \quad (5)$$

The reception of a word that is not in this set of codewords can be detected using the parity-check constraint equation (3). Suppose, for example, that the word $r = 101011$ is received from the channel. Substitution into Eq. (3) gives

$$Hr^T = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \quad (6)$$

which is nonzero, and so the word 101011 is not a codeword of our code.

To go further and correct the error requires that the decoder determine the codeword most likely to have been sent. Since it is reasonable to assume that the number of errors will more likely be small rather than large, the required codeword is the one closest in *Hamming distance* to the received word. By comparison of the received word $r = 101011$ with each codeword in (5), the closest codeword is $c = 001011$, which is at Hamming distance 1 from r. The minimum distance of this code is 3, so a single bit error always results in a word closer to the codeword that was sent than any other codeword, and hence can always be corrected. In general, for a code with minimum distance $d_{\min}$, e bit errors can always be corrected by choosing the closest codeword whenever

$$e \leq \left\lfloor \frac{d_{\min} - 1}{2} \right\rfloor \quad (7)$$

where $\lfloor x \rfloor$ is the largest integer that is at most x.

Error correction by direct search is feasible only when the number of distinct codewords is small. For codes with thousands of bits in a codeword, it becomes far too computationally expensive to directly compare the received word with every codeword in the code, and numerous ingenious solutions have been proposed, including choosing codes that are cyclic or, as presented in this article, devising iterative methods to decode the received word.

2.2. Low-Density Codes

LDPC codes are parity-check codes with the requirement that H is low-density, so that the vast majority of entries are zero. A parity-check matrix is *regular* if each code bit is contained in a fixed number, w_c, of parity checks and each parity-check equation contains a fixed number, w_r, of code bits. If an LDPC code is described by a regular parity-check matrix it is called a (w_c, w_r)-*regular LDPC* code; otherwise it is an *irregular LDPC* code.

Importantly, an error correction code can be described by more than one *parity-check matrix*, where H is a valid parity-check matrix for a code, provided (3) holds for all codewords in the code. Two parity-check matrices for the same code need not even have the same number of rows; what is required is that the rank over GF(2) of both be the same, since the number of message bits, k, in a binary code is

$$k = n - \text{rank}_2(H) \quad (8)$$

where $\text{rank}_2(H)$ is the number of rows in H that are linearly dependent over GF(2). To illustrate, we give a regular parity-check matrix for the code of (2) with $w_c = 2$, $w_r = 3$, and $\text{rank}_2(H) = 3$, which satisfies (3)

$$H = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix} \quad (9)$$

A *Tanner graph* is a graphical representation of H that facilitates iterative decoding of the code. The Tanner

graph consists of two sets of vertices: *n bit vertices* (or *bit nodes*) and *m parity-check vertices* (or *check nodes*), where there is a parity-check vertex for every parity-check equation in H and a bit vertex for every codeword bit. Each parity-check vertex is connected by an edge to the bit vertices corresponding to the code bits included in that parity-check equation. The Tanner graph of the parity-check matrix (9) is shown in Fig. 2. As the number of edges leaving the bit vertices must equal the number of edges leaving the parity-check vertices it follows that for a regular code:

$$m \cdot w_r = n \cdot w_c \tag{10}$$

A *cycle* in a Tanner graph is a sequence of connected vertices that start and end at the same vertex in the graph, and that contain other vertices no more than once. The length of a cycle is the number of edges it contains, and the *girth* of a graph is the size of its smallest cycle. A cycle of size 6 is shown in bold in Fig. 2.

Traditionally, the parity-check matrices of LDPC codes have been defined pseudorandomly subject to the requirement that H be sparse, and code construction of binary LDPC codes involves randomly assigning a small number of the values in an all-zero matrix to be 1. The lack of any obvious algebraic structure in randomly constructed LDPC codes sets them apart from traditional parity-check codes. The properties and performance of LDPC codes are often considered in terms of the ensemble performance of all possible codes with a specified structure (e.g., a certain node degree distribution), reminiscent of the methods used by Shannon in proving his noisy channel coding theorem. More recent research has considered the design of LDPC codes with specific properties, such as large girth, and we describe in Sections 5 and 6 methods to design LDPC codes. For sum–product decoding, however, no additional structure beyond a sparse parity-check matrix is required, and in the following two sections we present decoding algorithms requiring only the existence of a sparse H.

3. ITERATIVE DECODING

To illustrate the process of iterative decoding, a *bit-flipping* algorithm is presented, based on an initial hard decision (0 or 1) assessment of each received bit. An essential part of iterative decoding is the passing of messages between the nodes of the Tanner graph of the code. For the bit-flipping algorithm, the messages are simple; a bit node sends a message to each of the check nodes to which it is connected, declaring whether it is a 1 or a 0, and each check node sends a message to each of the bit nodes to which it is connected, declaring whether the parity check is satisfied. The sum–product algorithm for LDPC codes operates similarly but with more complicated messages.

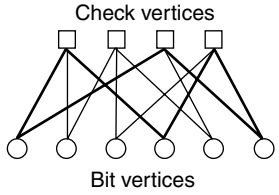


Figure 2. Tanner graph representation of the parity-check matrix in (9). A 6-cycle is shown in bold.

The bit-flipping decoding algorithm is as follows:

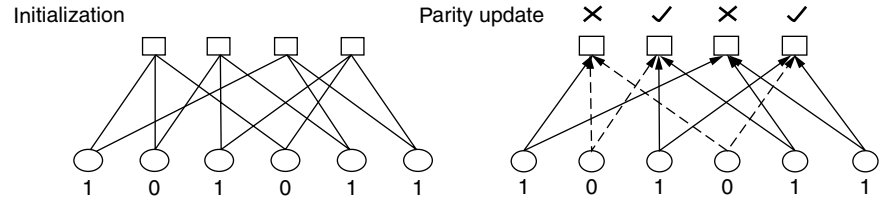
Step 1. Initialization. Each bit node is assigned the bit value received from the channel, and sends messages to the check nodes to which it is connected indicating this value.

Step 2. Parity update. Using the messages from the bit nodes, each check node calculates whether its parity-check equation is satisfied. If all parity-check equations are satisfied, the algorithm terminates; otherwise each check node sends messages to the bit nodes to which it is connected indicating whether the parity-check equation is satisfied.

Step 3. Bit update. If the majority of the messages received by each bit node are "not satisfied," the bit node flips its current value; otherwise the value is retained. If the maximum number of allowed iterations is reached, the algorithm terminates and a failure to converge is reported; otherwise each bit node sends new messages to the check nodes to which it is connected, indicating its value, and the algorithm returns to step 2.

To illustrate the operation of the bit-flipping decoder, we take the code of (9) and again assume that the codeword $c = 001011$ is sent, and the word $r = 101011$ is received from the channel. The steps required to decode this received word are shown in Fig. 3. In step 1 the bit values are initialized to be 1, 0, 1, 0, 1, and 1, respectively, and messages are sent to the check nodes indicating these values. In step 2 each parity-check equation is satisfied only if an even number of the bits included in the parity-check equation are 1. For the first and third check nodes this is not the case, and so they send "not satisfied" messages to the bits to which they are connected. In step 3 the first bit has the majority of its messages indicating "not satisfied" and so flips its value from 1 to 0. Step 2 is repeated, and since now all four parity-check equations are satisfied, the algorithm halts and returns $c = 001011$ as the decoded codeword. The received word has therefore been correctly decoded without requiring an explicit search over all possible codewords.

The existence of cycles in the Tanner graph of a code reduces the effectiveness of the iterative decoding process. To illustrate the detrimental effect of a 4-cycle, we adjust the code of the previous example to obtain the new code shown in Fig. 4. A valid codeword for this code is 001001, but again we assume that the first bit is corrupted, so that $r = 101001$ is received from the channel. The steps of the bit-flipping algorithm for this received word are shown in Fig. 4. In step 1 the initial bit values are 1, 0, 1, 0, 0, and 1, respectively, and messages are sent to the check nodes indicating these values. Step 2 reveals that the first and second parity-check equations are not satisfied. In step 3 both the first and second bits have the majority of their messages indicating "not satisfied," and so both flip their bit values. When step 2 is repeated, we see that the first and second parity-check equations are again not satisfied.



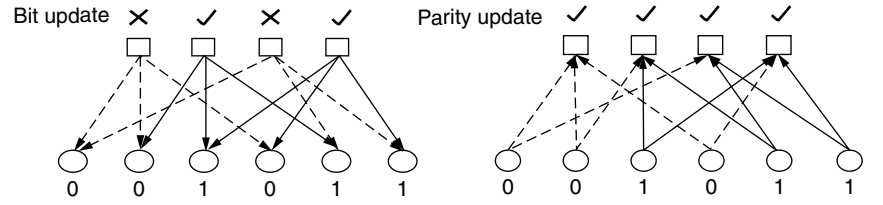
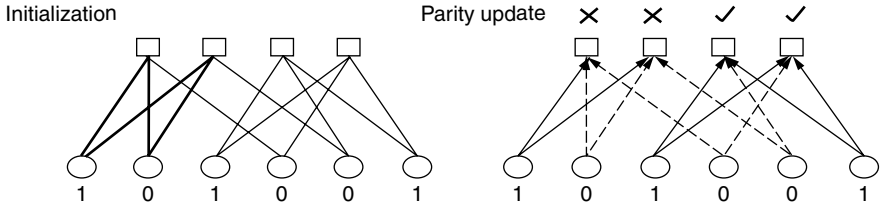


Figure 3. Bit-flipping decoding of the received word $r = 101011$. Each diagram indicates the decision made at each step of the decoding algorithm based on the messages from the previous step. A cross ($\times$) represents that the parity check is not satisfied, while a tick ($\surd$) indicates that it is satisfied. For the messages, a dashed arrow corresponds to the messages "bit = 0" or "check not satisfied," while a solid arrow corresponds to "bit = 1" or "check satisfied."



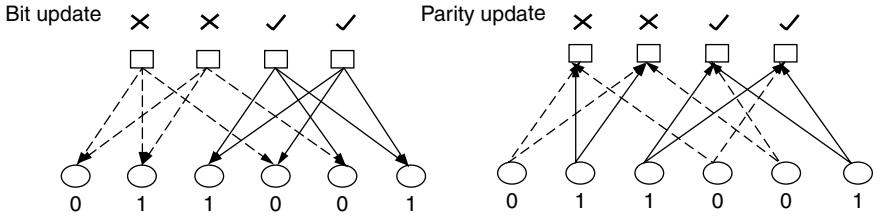


Figure 4. Bit-flipping decoding of the received word $r = 101001$. A 4-cycle is shown in bold in the first diagram.

Further iterations at this point simply cause the first 2 bits to flip their values in such a way that one of them is always incorrect; the algorithm fails to converge. As a result of the 4-cycle, each of the first two codeword bits is involved in the same two parity-check equations, and so when neither of the parity-check equations is satisfied, it is not possible to determine which bit is causing the error.

4. SUM–PRODUCT DECODING

The sum–product decoding algorithm, also called *belief propagation* decoding, was first introduced by Gallager in his 1962 thesis, where he applied it to the decoding of pseudorandomly constructed LDPC codes. For block lengths of 10^7, highly optimized irregular LDPC codes decoded with the sum-product algorithm are now known to be capable of approaching the Shannon limit to within hundredths of a decibel on the binary input additive white Gaussian noise (AWGN) channel. In the early 1960s, however, limited computing resources prevented Gallager from demonstrating the capabilities of iteratively decoded LDPC codes for blocklengths longer than $\sim$500, and for over 30 years his work was ignored by only a handful of researchers. It was only rediscovered by several researchers in the wake of Turbo decoding [4], which has subsequently been recognized as an instance of the sum–product algorithm.

The sum–product algorithm can be regarded as being similar to the bit-flipping algorithm described in the previous section, but with the messages representing each decision (check met, or bit value equal to 1) now probabilistic values represented by loglikelihood ratios. Whereas with bit-flipping decoding an initial hard decision is made on the signal from the channel, what is actually received is a string of real values where the sign of the received value represents a binary decision — 0 if positive and 1 if negative — and the magnitude of the received value is a measure of the confidence in that decision. A shortcoming of using only hard decisions when decoding is that the information relating to the confidence of the signal, the soft information, is discarded. Soft-decision decoders, such as the sum–product decoder, make use of the soft received information, together with knowledge of the channel properties, to obtain probabilistic expressions for the transmitted signal.

For a binary signal, if p is the probability of a 1, then $1 - p$ is the probability of a 0 that is represented as a

loglikelihood ratio (LLR) by

$$\mathrm{LLR}(p) = \log_e\left(\frac{1-p}{p}\right) \tag{11}$$

The sign of LLR(p) is the hard decision, and the magnitude |LLR(p)| is the reliability of this decision. One benefit of the logarithmic representation of probabilities is that whereas probabilities need to be multiplied, loglikelihood ratios need only be added, reducing implementation complexity.

The aim of sum–product decoding is to compute the *a posteriori probability* (APP) for each codeword bit, $P_i = P\{c_i = 1 \mid N\}$, which is the probability that the ith codeword bit is a 1 conditional on the event N that all parity-check constraints are satisfied. The *intrinsic* or *a priori probability*, P_i^{int}, is the original bit probability independent of knowledge of the code constraints, and the *extrinsic* probability P_i^{ext} represents what has been learnt from the event N.

The sum–product algorithm iteratively computes an approximation of the APP value for each code bit. The approximations are exact if the code is cycle-free. Extrinsic information gained from the parity-check constraints in one iteration is used as a priori information for the subsequent iteration. The extrinsic bit information obtained from a parity-check constraint is independent of the a priori value for that bit at the start of the iteration. The extrinsic information provided in subsequent iterations remains independent of the original a priori probability until that information is returned via a cycle.

To compute the extrinsic probability of a codeword bit i from the jth parity-check equation, we determine the probability that the parity-check equation is satisfied if bit i is assumed to be a 1, which is the probability that an odd number of the other codeword bits are a 1:

$$P_{i,j} = \frac{1}{2} + \frac{1}{2}\prod_{i' \in B_j, i' \neq i}(1 - 2P_{i'}^{\mathrm{int}}) \tag{12}$$

The notation B_j represents the set of column locations of the bits in the jth parity-check equation of the code considered. Similarly, A_i is the set of row locations of the parity-check equations which check on the ith bit of the code. To put (12) into loglikelihood notation we note that

$$\tanh\left(\frac{1}{2}\log_e\left(\frac{1-p}{p}\right)\right) = 1 - 2p$$

to give

$$\mathrm{LLR}(P_{i,j}^{\mathrm{ext}}) = \log_e\left(\frac{1 + \prod_{i' \in B_j, i' \neq i}\tanh(\mathrm{LLR}(P_{i'}^{\mathrm{int}})/2)}{1 - \prod_{i' \in B_j, i' \neq i}\tanh(\mathrm{LLR}(P_{i'}^{\mathrm{int}})/2)}\right)$$

The LLR of the estimated APP of the ith bit at each iteration is then simply

$$\mathrm{LLR}(P_i) = \mathrm{LLR}(P_i^{\mathrm{int}}) + \sum_{j \in A_i}\mathrm{LLR}(P_{i,j}^{\mathrm{ext}})$$

The sum–product algorithm is as follows:

Step 1. Initialization. The initial message sent from bit node i to the check node j is the LLR of the (soft) received signal y_i given knowledge of the channel properties. For an AWGN channel with signal-to-noise ratio E_b/N_0, this is

$$L_{i,j} = R_i = 4y_i\frac{E_b}{N_0} \tag{13}$$

Step 2. Check to bit. The extrinsic message from check node j to bit node i is the probability that parity check j is satisfied if bit i is assumed to be a 1 expressed as an LLR:

$$E_{i,j} = \log_e\left(\frac{1 + \prod_{i' \in B_j, i' \neq i}\tanh(L_{i'j}/2)}{1 - \prod_{i' \in B_j, i' \neq i}\tanh(L_{i'j}/2)}\right) \tag{14}$$

Step 3. Codeword test. The combined LLR is the sum of the extrinsic LLRs and the original LLR calculated in step 1:

$$L_i = \sum_{j \in A_i} E_{i,j} + R_i \tag{15}$$

For each bit a hard decision is made:

$$z_i = \begin{cases} 1, & L_i \leq 0 \\ 0, & L_i > 0 \end{cases}$$

If $z = [z_1, \ldots, z_n]$ is a valid codeword ($Hz^T = 0$), or if the maximum number of allowed iterations have been completed, the algorithm terminates.

Step 4. Bit to check. The message sent by each bit node to the check nodes to which it is connected is similar to (15), except that bit i sends to check node j a LLR calculated without using the information from check node j:

$$L_{i,j} = \sum_{j' \in A_i, j' \neq j} E_{i,j'} + R_i \tag{16}$$

Return to step 2.

The application of Eqs. (14) and (16) to the code in (9) is demonstrated in Fig. 5. The extrinsic information passed from a check node to a bit node is independent of the probability value for that bit. The extrinsic information from the check nodes is then used as a priori information for the bit nodes in the subsequent iteration.

To illustrate the power of sum–product decoding, we revisit the example of Fig. 3, where the codeword sent is 0 0 1 0 1 1. Suppose that the channel is AWGN with $E_b/N_0 = 1.25$ and the received signal is $y = -0.1\ \ 0.5\ \ -0.8\ \ 1.0\ \ -0.7\ \ 0.5$. There are now two bits in error if the hard decision of the signal is considered: bits 1 and 6. Figure 6 illustrates the operation of the sum–product decoding algorithm, as described in Eqs. (13)–(16), to decode this received signal which terminates in three iterations. The existence of an exact termination rule for the sum–product algorithm has two important benefits: (1) a failure to converge is always

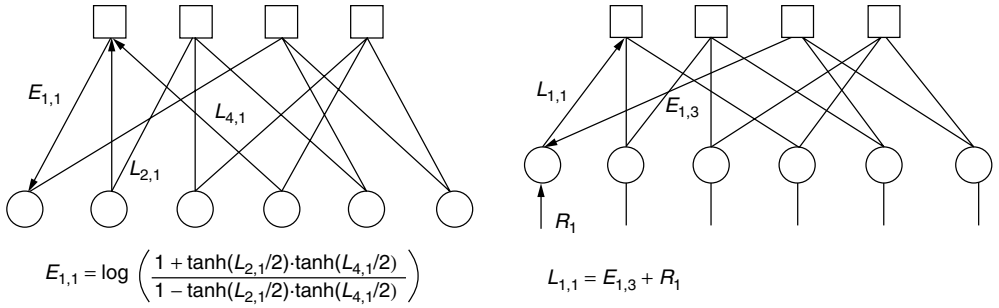


Figure 5. An example of the messages for sum–product decoding. Calculation of the extrinsic message sent to bit 1 depends on messages from bits 2 and 4 but not from bit 1. Similarly, the message sent to check 1 is independent of the message just received from it.

Iteration 1

$R = [\,-0.5000 \quad 2.5000 \quad -4.0000 \quad 5.0000 \quad -3.5000 \quad 2.5000\,]$

$[\,1\ 0\ 1\ 0\ 1\ 0\,]$ as a hard decision

$$E = \begin{bmatrix} 2.4217 & -0.4930 & . & -0.4217 & . & . \\ . & 3.0265 & -2.1892 & . & -2.3001 & . \\ -2.1892 & . & . & . & -0.4217 & 0.4696 \\ . & . & 2.4217 & -2.3001 & . & -3.6869 \end{bmatrix}$$

$L = [\,-0.2676 \quad 5.0334 \quad -3.7676 \quad 2.2783 \quad -6.2217 \quad -0.7173\,]$

$z = [\,1\ 0\ 1\ 0\ 1\ 1\,]$

$Hz^T = [\,1\ 0\ 1\ 0\,]^T \Rightarrow$ Continue

$$L = \begin{bmatrix} -2.6892 & 5.5265 & . & 2.6999 & . & . \\ . & 2.0070 & -1.5783 & . & -3.9217 & . \\ 1.9217 & . & . & . & -5.8001 & -1.1869 \\ . & . & -6.1892 & 4.5783 & . & 2.9696 \end{bmatrix}$$

Iteration 2

$$E = \begin{bmatrix} 2.6426 & -2.0060 & . & -2.6326 & . & . \\ . & 1.4907 & -1.8721 & . & -1.1041 & . \\ 1.1779 & . & . & . & -0.8388 & -1.9016 \\ . & . & 2.7877 & -2.9305 & . & -4.3963 \end{bmatrix}$$

$L = [\quad 3.3206 \quad 1.9848 \quad -3.0845 \quad -0.5630 \quad -5.4429 \quad -3.7979\,]$

$z = [\,0\ 0\ 1\ 1\ 1\ 1\,]$

$Hz^T = [\,1\ 0\ 0\ 1\,]^T \Rightarrow$ Continue

$$L = \begin{bmatrix} 0.6779 & 3.9907 & . & 2.0695 & . & . \\ . & 0.4940 & -1.2123 & . & -4.3388 & . \\ 2.1426 & . & . & . & -4.6041 & -1.8963 \\ . & . & -5.8721 & 2.3674 & . & 0.5984 \end{bmatrix}$$

Iteration 3

$$E = \begin{bmatrix} 1.9352 & 0.5180 & . & 0.6515 & . & . \\ . & 1.1733 & -0.4808 & . & -0.2637 & . \\ 1.8332 & . & . & . & -1.3362 & -2.0620 \\ . & . & 0.4912 & -0.5948 & . & -2.3381 \end{bmatrix}$$

$L = [\quad 3.2684 \quad 4.1912 \quad -3.9896 \quad 5.0567 \quad -5.0999 \quad -1.9001\,]$

$z = [\,0\ 0\ 1\ 0\ 1\ 1\,]$

$Hz^T = [\,0\ 0\ 0\ 0\,]^T \Rightarrow$ Terminate

Figure 6. Operation of sum–product decoding with the code from (9) when the codeword [001011] is sent through an AWGN channel with $E_b/N_0 = 1.25$ and the vector $[-0.1\ 0.5\ -0.8\ 1.0\ -0.7\ 0.5]$ is received. The sum–product decoder converges to the correct codeword after three iterations.

detected, and (2) additional iterations are avoided once a solution has been found.

There are variations to the sum–product algorithm presented here. The *min-sum* algorithm, for example, simplifies the calculation of (14) by recognizing that the term corresponding to the smallest $L_{i'j}$ dominates the product term, and so the product can be approximated by a minimum; the resulting algorithm thus requires calculation of only minimums and additions. An alternative approach, designed to bridge the gap between the error performance of sum-product decoding and that of maximum-likelihood (ML) decoding, finishes each iteration of sum–product decoding with ordered statistic decoding, with the algorithm terminating when a specified number of iterations have returned the same codeword [5].

5. CODES, GRAPHS, AND CYCLES

The relationship between LDPC codes and their decoding is closely associated with the graph-based representations of the codes. The most obvious example of this is the link between the existence of cycles in the Tanner graph of the code to both the analysis and performance of sum–product decoding of the code. In his work, Gallager used a graphical representation of the bit and parity-check sets of regular LDPC codes, to describe the application of iterative APP decoding. The systematic study of codes on graphs, however, is due largely to Tanner, who, in 1981, extended the single parity-check constraints of Gallager's LDPC codes to arbitrary linear code constraints, foresaw the advantages for very large-scale integration (VLSI) implementations of iterative decoders, and formalized the use of bipartite graphs for describing families of codes [6]. In so doing, Tanner also founded the topic of algebraic methods for constructing graphs suitable for sum–product decoding.

By proving the convergence of the sum–product algorithm for codes whose graphs are free of cycles, Tanner was also the first to formally recognize the importance of cycle-free graphs in the context of iterative decoding. The effect of cycles on the practical performance of LDPC codes was demonstrated by simulation experiments when LDPC codes were rediscovered by MacKay and Neal [7] (among others) in the mid-1990s, and the beneficial effects of using graphs free of short cycles were shown [8]. Given the detrimental effects of cycles on the convergence of iterative decoders, it is natural to seek strong codes whose Tanner graphs are free of cycles. An important negative result in this direction was established by Etzion et al. [9], who showed that for linear codes of rate $k/n \geq 0.5$, which can be represented by a Tanner graph without cycles, the minimum distance is at most 2.

As the existence of cycles in a graph makes analysis of the decoding algorithm difficult, most analyses consider the asymptotic performance of iterative decoding on graphs with asymptotically unbounded girth. This analysis provides thresholds to the performance of LDPC codes with sum–product decoding. As we will see in the following section, this process can be used to select LDPC code properties that improve the threshold values, a process that works well even though the resulting codes contain cycles.

To date very little analysis has been presented regarding the convergence of iterative decoding methods on graphs with cycles, and the majority of the work in this area can be found in Ref. 10. Gallager suggested that the dependencies introduced by cycles have a relatively minor effect and tend to cancel each other out somewhat. This "seems to work" philosophy has underlined the performance of sum–product decoding on graphs with cycles for much of the (short) history of the topic. It is only relatively recently that exact analysis on the expected performance of codes with cycles has emerged. Di et al. [11] use finite-length analysis to give the exact average bit and block error probabilities for any regular ensemble of LDPC codes over the binary erasure channel when decoded iteratively; however, there is as yet no such analysis for irregular codes or more general channel models.

Besides cycles, Sipser and Spielman [12] showed that the expansion of the graph is a significant factor in the application of iterative decoding. Using only a simple hard-decision decoding algorithm, they proved that a fixed fraction of errors in an LDPC code can be corrected in linear time provided that the Tanner graph of the code is a sufficiently good expander. That is, any subset S of bit vertices of size m or less is connected to at least $\in |S|$ constraint vertices, for some defined m and $\in$.

6. DESIGNING LDPC CODES

For the most part, LDPC codes are designed by first choosing the required blocklength and node degree distributions, then pseudorandomly constructing a parity-check matrix, or graph, with these properties. A generator matrix for the code can then be found using Gaussian elimination [8]. Gallager, for example, considered the ensemble of all (w_r, w_c)-regular matrices with rows divided into w_c submatrices, where the first contain w_r copies of the identity matrix and subsequent submatrices are random column permutations of the first. Using ensembles of matrices defined in this way, Gallager was able to find the maximum crossover probability of the binary symmetric channel (BSC) for which LDPC codes could be used to transmit information reliably using a simple hard-decision decoding algorithm.

Luby et al.extended the class of LDPC ensembles to those with irregular node degrees and showed that irregular codes are capable of outperforming regular codes [13]. In extending Gallager's analysis to irregular ensembles, Luby et al. introduced tools based on linear programming for designing irregular code ensembles for which the maximum allowed crossover probability of the binary symmetric channel is optimized [14]. Resulting from this work are the "tornado codes," a family of codes that approach the capacity of the erasure channel and can be encoded and decoded in linear time.

Richardson and Urbanke extended the work of Luby et al. to any binary input memoryless channel and to soft-decision message-passing decoding [15]. They determined the capacity of message-passing decoders applied to LDPC code ensembles by a method called *density evolution*.

For sum–product decoding density evolution makes it possible to determine the corresponding capacity to any degree of accuracy and hence determine the ensemble with node degree distribution that gives the best capacity. Once a code ensemble has been chosen a code from that ensemble is realized pseudorandomly. By carefully choosing a code from an optimized ensemble, Chung et al. have demonstrated the best performance to date of LDPC codes in terms of approaching the Shannon limit [16].

A more recent development in the design of LDPC codes is the introduction of algebraic LDPC codes, the most promising of which are the finite-geometry codes proposed by Lucas et al. [17], which are cyclic and described by sparse 4-cycle free graphs. An important outcome of this work with finite-geometry codes was the demonstration that highly redundant parity-check matrices can lead to very good iterative decoding performance without the need for very long blocklengths. Although the probability of a random graph having a highly redundant parity-check matrix is vanishingly small, the field of *combinatorial designs* offers a rich source of algebraic constructions for matrices that are both sparse and redundant. In particular, there has been much interest in balanced incomplete block designs (BIBDs) to produce sparse matrices for LDPC codes that are 4-cycle-free. For codes with greater girth, generalized quadrangle designs give the maximum possible girth for a graph with given diameter [18]. Both generalized quadrangles and BIBDs are subsets of the more general class of combinatorial structures called *partial geometries*, a possible source of further good algebraic LDPC codes [19].

In comparison with more traditional forms of error-correcting codes, the minimum distance of LDPC codes plays a substantially reduced role. There are two reasons for this: (1) the lack of any obvious algebraic structure in pseudorandomly constructed LDPC codes makes the calculation of minimum distance infeasible for long codes, and most analyses focus on the average distance function for an ensemble of LDPC codes; and (2) the absence of conspicuous flattening of the bit-error-rate (BER) curve at moderate to high signal-to-noise ratios (the "error floor") strongly suggests that minimum distance properties are simply not as important for LDPC codes as for traditional codes. Indeed, it has been established that to achieve capacity on the binary erasure channel when using irregular LDPC codes, the codes cannot have large minimum distances [20].

7. CONNECTIONS AND FUTURE DIRECTIONS

Following the rediscovery of Gallager's iterative LDPC decoding algorithm in the mid-1990s, the notion of an iterative algorithm operating on a graph has been generalized and is now capable of unifying a wide range of apparently different algorithms from the domains of digital communications, signal processing, and even artificial intelligence. An important generalization of Tanner graphs was presented by Wiberg in his 1996 Ph.D. thesis [21]. Wiberg introduced *state variables* into the graphical framework, thereby establishing a connection between codes on graphs and the trellis complexity of codes, and was the first to observe that on cycle-free graphs, the sum–product (respectively, min-sum) algorithm performs APP (respectively, ML) decoding.

In an even more general setting, the role of the Tanner graph is taken by a *factor graph* [22]. Central to the unification of message-passing algorithms via factor graphs is the recognition that many computationally efficient signal processing algorithms exploit the manner in which a global cost function acting on many variables can be factorized into the product of simpler local functions, each of which operates on a subset of the variables. In this setting a (bipartite) factor graph encodes the factorization of the global cost function, with each local function node connected by edges only to those variable nodes associated with its arguments.

The sum–product algorithm operating on a factor graph uses message passing to solve the *marginalize product-of-functions* (MPF) problem which lies at the heart of many signal processing problems. In addition to the iterative decoding of LDPC codes, specific instances of the sum–product algorithm operating on suitably defined factor graphs include the forward/backward algorithm (also known as the *BCJR* (Bahl–Cocke–Jelinek–Raviv) *algorithm* [23] or *APP decoding algorithm*), the Viterbi algorithm, the Kalman filter, Pearl's belief propagation algorithm for Bayesian networks, and the iterative decoding of "*Turbo codes*," or parallel concatenated convolutional codes.

For high-performance applications, LDPC codes are naturally seen as competitors to Turbo codes. LDPC codes are capable of outperforming Turbo codes for blocklengths greater than $\sim 10^5$, and the error floors of LDPC codes at BERs below $\sim 10^{-5}$ are typically much less pronounced than those of Turbo codes. Moreover, the inherent parallelism of the sum–product decoding algorithm is more readily exploited with LDPC codes than their Turbo counterparts, where block interleavers pose formidable challenges to achieving high throughput [24]. Despite these impressive advantages, LDPC codes lag behind Turbo codes in real-world applications. The exceptional simulation performance of the original Turbo codes [4,25] generated intense interest in these codes, and variants of them were subsequently incorporated into proposals for third-generation (3G) wireless systems such as the Third Generation Partnership Project (3GPP), a global consortium of standards-setting organizations [26]. Whatever performance advantages of very long LDPC codes over Turbo codes there may be, the invention of Turbo codes some 3 years prior to the (re)discovery of LDPC codes has given them a distinct advantage in wireless communications, where blocklengths of at most several thousand are typical, and where compliance with global standards is paramount.

One serious shortcoming of LDPC codes is their potentially high *encoding* complexity, which is in general

quadratic in the blocklength, and compares poorly with the linear time encoding of Turbo codes. Finding computationally efficient encoders is therefore critical for LDPC codes to be considered as serious contenders for replacing Turbo codes in future generations of forward error correction devices. Several approaches have been suggested, including the manipulation of the parity-check matrix to establish that while the complexity is, strictly speaking, quadratic, the actual number of encoding operations grows essentially linearly with blocklength. For some irregular LDPC codes whose degree distributions have been optimized to allow transmission near to capacity, the encoding complexity can be shown to be truly linear in blocklength [27]. A very different approach to the encoding complexity problem is to employ cyclic, or quasicyclic, codes as LDPC codes, as encoding can be achieved in linear time using simple feedback shift registers [28].

While addressing encoding complexity is driven by applications, two issues seem likely to dominate future theoretical investigations of LDPC codes. The first of these is to characterize the performance of LDPC codes with ML decoding and thus to assess how much loss in performance is due to the structure of the codes, and how much is due to the suboptimum iterative decoding algorithm. The second, and related, issue is to rigorously deal with the decoding of codes on graphs with cycles. Most analyses to date have assumed that the graphs are effectively cycle-free. What is not yet fully understood is just why the sum–product decoder performs as well as it does with LDPC codes having cycles.

BIOGRAPHIES

Sarah J. Johnson was born in 1977, and received the B.E. degree in electrical engineering in 2000 (Hons I and University Medal) from the University of Newcastle, Australia. She is presently a candidate for the Ph.D. degree in electrical engineering at the University of Newcastle, Australia, where her research interests include low-density parity-check codes, and iterative decoding algorithms.

Steven R. Weller was born in Sydney, Australia, in 1965. He received the B.E. (Hons I) degree in computer engineering in 1988, the M.E. degree in electrical engineering in 1992, and the Ph.D. degree in electrical engineering in 1994, all from the University of Newcastle, Australia. From April 1994 to July 1997 he was a lecturer in the Department of Electrical and Electronic Engineering at the University of Melbourne, Australia, and was a member of the Centre for Sensor Signal and Information Processing (CSSIP). Since July 1997 he has been at the University of Newcastle, Australia, where he is currently a Senior Lecturer in the School of Electrical Engineering and Computer Science, and a member of the Centre for Integrated Dynamics and Control (CIDAC). His research interests include low-density parity-check codes, iterative decoding algorithms, space time-coded communications, and combinatorics.

BIBLIOGRAPHY

1. C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* **27**: 379–423, 623–656 (July-Oct. 1948).
2. R. G. Gallager, Low-density parity-check codes, *IRE Trans. Inform. Theory* **IT-8**(1): 21–28 (Jan. 1962).
3. R. G. Gallager, *Low-Density Parity-Check Codes*, MIT Press, Cambridge, MA, 1963.
4. C. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error-correcting coding and decoding: Turbo-codes, *Proc. IEEE Int. Conf. Communications* (*ICC'93*), Geneva, Switzerland, May 1993, pp. 1064–1070.
5. M. P. C. Fossorier, Iterative reliability-based decoding of low-density parity check codes, *IEEE J. Select. Areas Commun.* **19**(5): 908–917 (May 2001).
6. R. M. Tanner, A recursive approach to low complexity codes, *IEEE Trans. Inform. Theory* **IT-27**(5): 533–547 (Sept. 1981).
7. D. J. C. MacKay and R. M. Neal, Near Shannon limit performance of low density parity check codes, *Electron. Lett.* **32**(18): 1645–1646 (March 1996); reprinted in *Electron. Lett.* **33**(6): 457–458 (March 1997).
8. D. J. C. MacKay, Good error-correcting codes based on very sparse matrices, *IEEE Trans. Inform. Theory* **45**(2): 399–431 (March 1999).
9. T. Etzion, A. Trachtenberg, and A. Vardy, Which codes have cycle-free Tanner graphs? *IEEE Trans. Inform. Theory* **45**(6): 2173–2181 (Sept. 1999).
10. *IEEE Trans. Inform. Theory* (Special issue on Codes on Graphs and Iterative Algorithms) **47**(2) (Feb. 2001).
11. C. Di etal, Finite-length analysis of low-density parity-check codes on the binary erasure channel, *IEEE Trans. Inform. Theory* **48**(6): 1570–1579 (June 2002).
12. M. Sipser and D. A. Spielman, Expander codes, *IEEE Trans. Inform. Theory* **42**(6): 1710–1722 (Nov. 1996).
13. M. G. Luby, M. Mitzenmacher, M. A. Shokrollahi, and D. A. Spielman, Efficient erasure correcting codes, *IEEE Trans. Inform. Theory* **47**(2): 569–584 (Feb. 2001).
14. M. G. Luby, M. Mitzenmacher, M. A. Shokrollahi, and D. A. Spielman, Improved low-density parity-check codes using irregular graphs, *IEEE Trans. Inform. Theory* **47**(2): 585–598 (Feb. 2001).
15. T. J. Richardson and R. L. Urbanke, The capacity of low-density parity-check codes under message-passing decoding, *IEEE Trans. Inform. Theory* **47**(2): 599–618 (Feb. 2001).
16. S.-Y. Chung, G. D. Forney, Jr., T. J. Richardson, and R. Urbanke, On the design of low-density parity-check codes within 0.0045 dB of the Shannon limit, *IEEE Commun. Lett.* **5**(2): 58–60 (Feb. 2001).
17. R. Lucas, M. P. C. Fossorier, Y. Kou, and S. Lin, Iterative decoding of one-step majority logic decodable codes based on belief propagation, *IEEE Trans. Commun.* **48**(6): 931–937 (June 2000).
18. P. O. Vontobel and R. M. Tanner, Construction of codes based on finite generalized quadrangles for iterative decoding, *Proc. IEEE Int. Symp. Information Theory*, Washington, DC, June 24–29, 2001, p. 223.
19. S. J. Johnson and S. R. Weller, Codes for iterative decoding from partial geometries, *Proc. IEEE Int. Symp. Information Theory*, Lausanne, Switzerland, June 30–July 5, 2002, p. 310.

20. C. Di, T. J. Richardson, and R. L. Urbanke, Weight distributions: How deviant can you be? *Proc. IEEE Int. Symp. Information Theory*, Washington, DC, June 24–29, 2001, p. 50.

21. N. Wiberg, *Codes and Decoding on General Graphs*, Ph.D. thesis, Dept. Electrical Engineering, Linköping Univ., Sweden, 1996.

22. F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, Factor graphs and the sum-product algorithm, *IEEE Trans. Inform. Theory* **47**(2): 498–519 (Feb. 2001).

23. L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* **IT-20**(2): 284–287 (March 1974).

24. A. J. Blanksby and C. J. Howland, A 690-mW 1-Gb/s 1024-b, rate-1/2 low-density parity-check code decoder, *IEEE J. Solid-State Circuits* **37**(3): 404–412 (March 2002).

25. C. Berrou and A. Glavieux, Near optimum error correcting coding and decoding: Turbo codes, *IEEE Trans. Commun.* **44**(10): 1261–1271 (Oct. 1996).

26. 3rd Generation Partnership Project (3GPP), *Technical Specification Group Radio Access Network; Multiplexing and Channel Coding* (*FDD*), 3GPP TS 25.212 V4.0.0 (2000-12) (online), *http://www.3gpp.org*.

27. T. J. Richardson and R. L. Urbanke, Efficient encoding of low-density parity-check codes, *IEEE Trans. Inform. Theory* **47**(2): 638–656 (Feb. 2001).

28. Y. Kou, S. Lin, and M. P. C. Fossorier, Low-density parity-check codes based on finite geometries: A rediscovery and new results, *IEEE Trans. Inform. Theory* **47**(7): 2711–2736 (Nov. 2001).

M

MAGNETIC STORAGE SYSTEMS

HEMANT K. THAPAR
LSI Logic Corporation
San Jose, California

1. INTRODUCTION

Data storage is an essential function within today's information delivery systems. While the communication of information focuses on the delivery "from here to there," the storage function is focused on the delivery of information "from now to then." The "now to then" may entail near real-time transactions, as in server-client systems, or a period of days or months, as in data archiving. Whatever the application, the demand for storage is exploding in computing, communication, consumer, and entertainment systems. Over 200 million magnetic hard disk drives and over 150 million optical drives, combining various forms of compact disk (CD) and digital video disk (DVD) drives, will be shipped worldwide in 2002. The market for magnetic hard disk drives alone is forecasted to grow to over 350 million drives by 2006. Using a conservative estimate of 100 Gbytes of average storage capacity per drive in that time frame, digital magnetic storage alone will support 35,000 petabytes (35×10^{18} bytes) of storage demand worldwide.

Storage devices can be broadly classified into two categories: solid-state and mechanical. Solid-state memories are based on semiconductor process technology, and they can be used as standalone components within a system, or integrated with other functions in monolithic form. Mechanical storage devices rely on the relative motion between a magnetic, optical, or hybrid (magneto-optic) transducer and an associated storage medium to store temporal signals as spatial patterns on the medium. They are complex standalone subsystems that are used to store large quantities of data in nonvolatile form; that is, the device power can be turned off while preserving the stored data. Solid-state memories consume less power during storage and retrieval of data and offer better mechanical reliability, but they are considerably more expensive than mechanical storage devices.

The trend in storage devices is similar to that for communication systems: digital storage is emerging as the technology of choice. While data storage is inherently digital, storage of ubiquitous analog sources of information, namely, audio and video, is being accomplished increasingly using digital techniques. Digital source coding methods, such as MPEG-x and its associated audio standard MP3, JPEG, and pulse code modulation (PCM), are used to convert the analog signals to digital bit sequences for the purposes of delivery and storage. The advantages of performance, cost, and flexibility are driving this trend. Digital storage offers the ability to maintain a low probability of error during repeated retrievals of the stored data. This advantage is similar to "regeneration" in digital communications. Sources with wide dynamic range signals, such as classical music, can be reproduced with low noise and distortion compared to analog storage. Digital audio tape (DAT), music CD, DVD, and PVR (personal video recorder) as a replacement for VCR are examples of the emerging trend to use digital storage. The cost of digital storage also continues to decline at a rapid rate because of the steady improvements in component technologies, as discussed later in the article, and the economies of scale associated with the mass personal computer market.

In order to meet the wide and varying demands of different applications, storage devices have become segmented on the basis of two factors: performance and cost per megabyte. Performance is measured in terms of the access time, which is defined as the average time spent to access the selected data. Figure 1 shows the segmentation of commonly used storage devices in various applications. Highly cost-sensitive applications, involving software distribution, consumer audio and video playback, use compact disk read only memory (CD-ROM) and DVD devices. These devices are based on the use of optical recording technology and are designed to support varying modes of storage, including read-only, write-once read-many (WORM), and erasable/rewriteable. Their access times typically are on the order of tens to hundreds of milliseconds, but the associated cost of the drive and media is very low. Data backup and archival storage rely largely on the lower cost, lower performing magnetic tape systems. Such systems have access times on the order of tens to hundreds of seconds, since they can only access the selected data sequentially. They, however, support very large volumetric densities (Mbytes/cu. ft.) at very low cost per megabyte. Near real-time transactional systems use the higher cost, higher performance hard disk drive (HDD) systems permitting direct access to the selected data in any arbitrary order. Such devices are based on the use of magnetic recording technology. Solid-state memory devices, with access times on the order of nanoseconds but much higher cost per megabyte, are used for real-time storage applications, such as caching, real-time data memory, and program control.

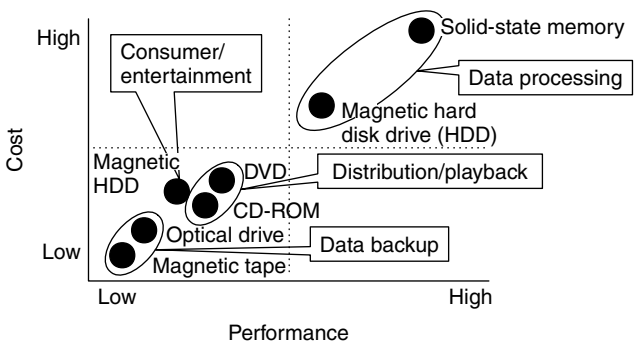


Figure 1. Storage devices segmentation.

Among mechanical storage devices, the magnetic HDD system has occupied a unique position in data storage since the introduction of IBM's RAMAC system in 1957 (see [15]). In order to keep this position unchallenged, the underlying magnetic component, signal processing, mechanical, and interface technologies have evolved at a very rapid pace to meet the growing demands of high-performance computing and peripheral devices. As a consequence of the rapid progress, digital magnetic storage technology is well poised to support the lower cost, lower performance segment previously serviced by optical or magnetic tape devices.[1] Magnetic HDD has already found its way in today's set-top boxes, allowing users to store in excess of 100 Gbytes of their favorite TV programs unattended. More applications in this area, dubbed "personal video recording," are likely to emerge as the much-vaunted convergence of communications, computing, entertainment, and mobility picks up pace. Similarly, IBM's one-inch diameter HDD is capable of storing 1 Gbyte of data in nonvolatile form at low cost. It is finding its way in digital cameras and other mobile applications. Except for removeability, digital magnetic recording offers everything that optical storage does, but with system attributes that include smaller, cheaper, denser, and faster.

This article provides an overview of digital magnetic storage based on the HDD system. Section 2 provides an overview of digital magnetic storage in HDD systems and the associated technology trends. Section 3 describes the digital magnetic recording channel, including the magnetic recording processes that underlie the generation of signals, noise, distortion, and interference during data retrieval. Section 4 describes the signal processing and coding techniques used in commercial HDD systems. Many of those techniques have their origins in data transmission and can be applied to other digital magnetic and optical storage channels with suitable modifications. Section 5 is devoted to concluding remarks.

2. HDD SYSTEM AND TRENDS

HDD systems are designed to deliver digital information "from now to then." Two processes are involved in such delivery: recording and retrieval.[2] The recording function, often referred to as "write," takes blocks of data (typically 4 kbits), appends control and synchronization information, and records it in the form of data sectors on the medium. The retrieval function, referred to as "read," processes the readback signal from the medium to deliver the recovered data. The two processes are similar to the transmit and the receive functions in data communication systems. While there are many similarities between data storage and communication, there are key differences that pertain to the error rate and synchronization requirements, which, in turn, have a bearing on the overall system design philosophy. Unlike data communications, HDD systems do not rely on "automatic request for retransmission" to recover from data errors; the retrieval process is designed to guarantee a prescribed worst-case bit error rate.[3] Similarly, since data recording involves mapping temporal data into spatial patterns, clock and data synchronization during retrieval must be fast and reliable to conserve the "real estate" on the medium. A great deal of effort is focused in HDD systems to minimize spatial overhead.

The HDD is a highly sophisticated electromechanical system. The mechanical assembly involves a slider mechanism holding a read/write head that flies about 10–20 nanometers (nm) over a rotating disk with speeds ranging from 3,600 rpm in the 1″ form factor (which refers to the disk diameter) HDD to 15,000 rpm in the 3.5″ form factor HDD. This head/media spacing places very stringent requirements on the disk surface in terms of uniformity, planarity, and defects. At the time of this writing, a typical drive will have 1 to 7 disks and 1 to 14 heads.

As shown in Fig. 2, data is stored on the disk in the form of spatial magnetic patterns along a *track*. The tracks are laid out as annuli of width W_t, which comprises the physical magnetic track width and a guard band between neighboring tracks. During data recording or retrieval, the head is positioned at a new track by moving the slider using servo control. To achieve accurate positioning, prewritten servo data patterns are interspersed along the tracks in the form of wedges, as shown in Fig. 2. These patterns are sensed during the head positioning process, which takes place in two major steps. First, the head seeks the track

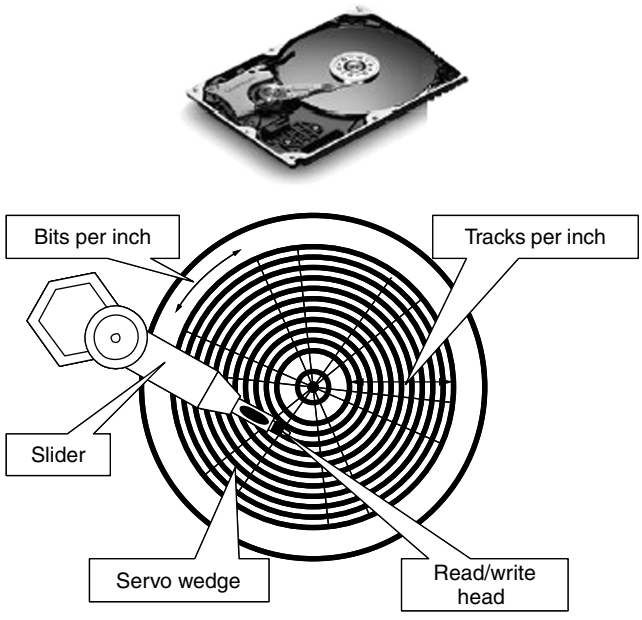


Figure 2. Data storage on hard disk drive.

[1] Even though Fig. 1 shows the cost for magnetic tape storage to be lower than HDD, the cost per megabyte for many HDD devices is comparable to that for tape systems.

[2] Accurate head positioning during recording and retrieval may be regarded as the third process for information delivery. It is very key to the operation of the HDD.

[3] Procedures that rely on the reread of the recorded data are also built into the HDD to recover from a rare error event, but their probability of use is minimized by design to maintain a high data throughput rate.

where the target data is to be located. The average amount of time required to seek the targeted track is called the *seek time*. The second step is to locate the data on the targeted track. The average time spent to locate the data is referred to as *latency*, which equals half the revolution period of the disk, since the target data location, on average, is halfway along the track from the initially positioned head. The sum of the seek time and latency defines the *access time*, which denotes the average time spent in going from one randomly selected location to another. Access time is a key measure of performance in data storage applications. Improvements in servo control algorithms, actuator design using lighter materials, and higher rotational speeds (e.g., 15000 rpm) are progressively reducing the access time in HDD. Today's products have access times ranging from 15 ms in IBM's 1″ Microdrive to below 6 ms in high performance server drives.

The number of bits stored per unit length of the track is referred to as *linear density*, measured in bits per inch (bpi). If the rotating speed of the disk is M revolutions per second, the data rate is R bits per second, and the track location is at radius r inches, then the linear density L is given by

$$L = \frac{R}{2\pi rM} \text{ bits/inch} \tag{1}$$

Note that the linear density grows towards infinity as r approaches 0. In practice, a nonzero inner radius, r_i, is selected to achieve a prescribed maximum linear density. Likewise, a prescribed outer radius, r_o, is used and the data storage is confined to the region between r_o and r_i. Based on capacity considerations (see Eq. (3) below), r_o is approximately equal to $2r_i$.

The number of tracks per unit length along the radial direction is referred to as the *track density*, measured in tracks per inch (tpi), and is given by

$$N_t = \frac{r_o - r_i}{W_t} \text{ tracks/inch} \tag{2}$$

The linear density is typically 8 to 15 times higher than the track density in commercially available products. The product of linear and track density defines the *areal density*, measured in bits per sq. inch. The storage capacity of a disk surface is the product of the areal density and the total surface area available for recording. Based on simple physical arguments, the capacity per surface, C, can be bounded as:

$$\frac{2\pi r_i(r_o - r_i)}{lW_t} \le C \le \frac{\pi({r_o}^2 - {r_i}^2)}{lW_t} \tag{3}$$

where W_t is the track width and l is the smallest bit cell length along the track. The upper bound assumes that bit cells of area lW_t are recorded over the entire disk surface.[4] Such a bound would be achieved if the linear velocity of the disk could be kept constant across all tracks. Since the linear velocity is radius-dependent (note $v = r\omega$, where ω is the angular velocity and r is the radius), it is impractical to keep it constant while supporting random access with low access time.[5] The lower bound is based on the use of constant data rate and rotational speed, wherein the number of bit cells at radius r_i, given by $2\pi r_i/l$, is kept constant across all tracks.

In practice, the capacity per surface lies between the two bounds. Instead of varying the rotational speed, the data rate is varied across the radii to effect better utilization of the disk surface. The disk surface is delineated into annular zones and the data rate is increased across the zones from the inner radius to the outer radius. This allows the zones along the outer radii to store more data, and hence yield higher storage capacity. The actual increase in capacity achieved from this so-called *zone-bit recording* scheme depends upon the number of zones and the linear density in each zone.

Areal density growth is key to increasing the capacity per disk surface. Indeed, the areal density has grown by a factor of 17 million since the introduction of the RAMAC drive in 1957 (see [13,14]). Figure 3 shows the areal density trends for commercial products and prototype demonstrations. At least two inflection points have occurred in the past decade. Since 1991, the rate of increase in areal density accelerated to 60% per year, and since 1997 this rate has further increased to 100% per year. Today's commercially available products store in excess of 50 Gbits/sq. inch, combining 80 ktpi in track density and 670 kbpi in linear density. Experimental prototype demonstrations exceeding 100 Gbits/sq. inch have been reported in the industry. The acceleration in 1991 of the annual growth rate was caused by the introduction of two key component technologies: magnetoresistive (MR) sensor for read heads and partial response maximum-likelihood (*PRML*) for read channels. The application of coding and PRML are discussed in more detail later in the chapter. The inflection point in 1997 was caused by the introduction of Giant MR (GMR) heads, which provide improved transducer sensitivity over their predecessors. Continual improvements in magnetic medium and signal processing technologies are also supporting this unprecedented growth rate in areal density.

The incredible growth in areal density has wrought similar trends in other figures-of-merit of interest in storage applications. Most importantly, the cost per megabyte decreases since the number of heads and disks required to achieve a prescribed capacity point decreases. Indeed, the cost per megabyte is declining at a rate of 40–50% annually, a rate currently higher than that for DRAM. Volumetric density also grows since smaller form factor disk drives,[6] with reduced head/disk count,

[4] In practice, the entire disk surface is not available for data storage. Some area is used to store servo data patterns for controlling the head positioning over the track as well as for storing overhead information related to defect skipping, calibration, etc.

[5] The linear velocity is kept constant across all tracks in compact audio players because the information is accessed sequentially from the inner radius to the outer radius and enough time is available to vary the rotational speed continually across the tracks.

[6] The 5.25 inches and larger form factors are all but obsolete now.

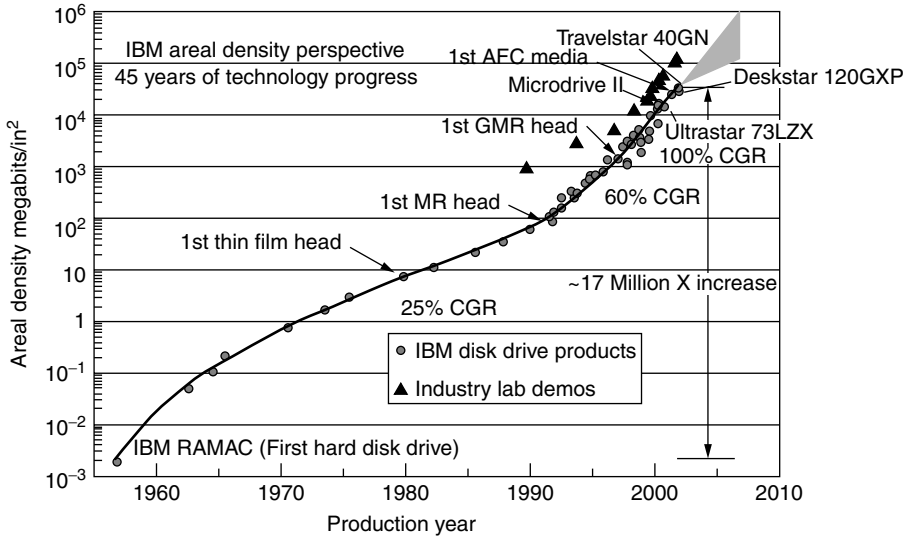


Figure 3. Area density trend in digital magnetic HDD.

achieve prescribed capacity points. The 3.5 inches and 2.5 inches form factors, which refer to the disk diameters, are mainstream today; the 1 inch is the emerging form factor. At the time of this writing, the 3.5 inches drive (with 1″ height), serving the high performance server market, stores approximately 80 Gbytes; the same form factor drive for desktop applications has a capacity exceeding 100 Gbytes. The 2.5 inches drive (with 1/2″ height) serving the mobile and notebook computers stores approximately 50 Gbytes. The 1.0 inch form factor (1/4″ height) stores 1 Gbyte. These same form factors are likely to double their storage capacity within a year, or support the same capacity with reduced number of heads and disks, and thus lowered cost per megabyte.

With smaller form factor and fewer disks, higher rotational speed and improved mechanical assembly can be achieved, thereby providing the means to reducing the access time. As the linear density grows with the areal density, the data transfer rate also increases (see Eq. (1)). The internal data transfer rates in today's disk drives is in the range of 400 Mbits/sec to over 1 Gbits/sec. It is growing at 30–40% annually. Together, the increasing transfer rates and decreasing access time are rendering HDD systems faster than before.

The above trends point to the following observation: magnetic disk drives are unequivocally becoming smaller, denser, faster, and cheaper. With these attributes, magnetic HDD is likely to become a viable storage device for such consumer applications as digital cameras, mobile communication devices, handheld computers, personal video recorders, and set-top boxes.

However, as the capacity per surface grows exponentially due to the increasing areal density, issues of reliability of the storage device become more acute. This trend has led to the development and proliferation redundant array of independent disks of (RAID) systems, which use redundancy within an array of disk drives to improve reliability and performance of the storage system. Tens of terabytes are aggregated in a RAID device with prescribed measures of data availability and reliability. Just as in error correction coding schemes, RAID devices are designed to reconstruct the stored data in the midst of a prescribed number of drive failures. Interconnected through data networks, these devices are used today to service the storage demands of the Internet and other information delivery systems with uncompromised availability.

The reader is referred to [13–15,28,30] for more detailed information on the trends for HDD systems.

3. DIGITAL MAGNETIC RECORDING CHANNEL

Digital magnetic recording is based on the elementary principles of electromagnetics wherein the temporal data signal to be recorded is converted into spatial patterns of magnets on a magnetic medium. It relies on the well-known *M-H* curve, shown in Fig. 4, which defines the switching behavior of the applied magnetic field, H, and the resulting remanent magnetization, M, of a magnetic material. Figure 4 shows that when the applied magnetic field exceeds the coercivity of the medium, H_c, the medium has remanent magnetization M_r. Likewise, when the field is reversed, the state of the magnetization is also reversed. Thus, the medium can be saturated into two

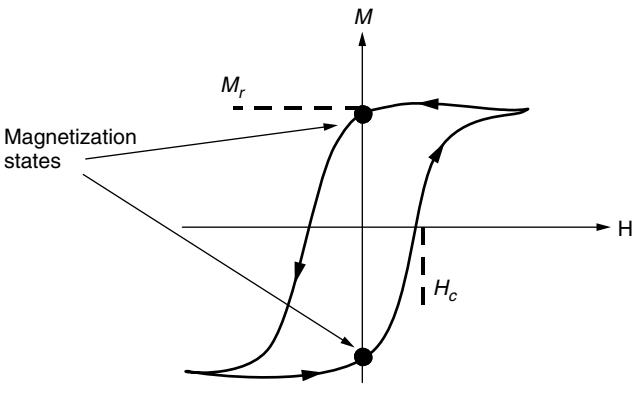


Figure 4. The *M-H* curve.

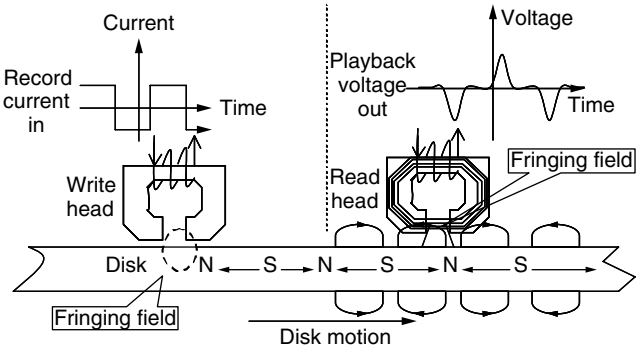


Figure 5. The digital magnetic recording and retrieval processes.

states, corresponding to binary 1s and 0s, by controlling the magnitude and direction of the applied field.

The recording and retrieval processes are illustrated in more detail in Fig. 5. As shown on the left side, the data sequence to be recorded is represented as a binary current waveform, which is applied to the write head with a gap. The flow of current in the windings of the write head generates a magnetic field within the head. The presence of the gap causes the field to fringe into a "bubble" and penetrate the magnetic recording medium. The field emanating from the gap can be decomposed into the longitudinal component (along the medium) and a perpendicular component. When the longitudinal component exceeds the coercivity of the medium, magnetic domains are created in accordance with field changes due to the current waveform. These domains are delineated in the figure by the symbols N and S, representing, respectively, the north and south poles of a magnet. Thus, temporal changes in the write current are mapped into spatial changes in the magnetic medium. Such a recording process is referred to as *longitudinal* saturation *recording*. With a different head-medium construction, one can use the perpendicular component of the head field to record the digital data. Such an approach, referred to as perpendicular recording, is regarded as a potential successor to longitudinal recording beyond 100 Gbits/sq. inch (see [28]).

During retrieval, the spatial magnetic patterns on the moving medium, represented by the alternating magnetic poles N and S, create fringing magnetic fields that are sensed by the read head, producing an analog voltage waveform that varies in accordance with the recorded patterns. Readback magnetic transducers are either inductive or magnetoresistive (MR). The inductive sensor produces a readback voltage that is proportional to the time derivative of the flux from the fringing field. The signal amplitude depends upon the rotational speed of the medium and the number of turns, N, in the winding of the inductive head [$V = -N(d\phi/dt)$]. MR heads use the MR stripe (or MR element), placed between two shields, to sense the flux from the external field. The change in flux causes a change in the resistance of the current-biased MR stripe, resulting in an output voltage that varies in accordance with the recorded magnetic patterns. The readback signal amplitude, unlike inductive heads, is independent of the rotational speed of the medium. Read heads based on MR-based sensors are ubiquitous in today's hard disk drive systems and remain major contributors to the accelerated areal density growth.

The preceding description is intended to capture the essence of the recording process. At the detailed level, the recording/retrieval processes are quite complex, especially as the bit cells continue to shrink. For example, the write current has finite rise times; the head fields do not switch instantaneously; magnetic transducers are frequency selective and often nonlinear; magnetic fields interact as transitions get closer due to the growth in linear density; and so on. All such factors cause nonidealities in the recording and replay processes, requiring sophisticated analysis, modeling, and experimental work to understand the signals, noise, interference, and distortion mechanisms. The reader is referred to Refs. 1, 4, 5, 30 for a more detailed treatment.

Based on the above description of the recording process, the recording channel characteristics are discussed below. The "channel" refers to the combined head and medium block that is responsible for generating the signals, noise, distortion, and interference mechanisms. Unlike communications channels, where the bandwidth and noise characteristics typically remain fixed after the spectral allocations are made, the digital magnetic recording channels continue to evolve and change. Signal and noise bandwidths get larger with the scaling of head/medium dimensions, and new noise phenomena arise as the bit cell dimensions shrink and new magnetic materials and transducers are introduced. The dynamic nature of the channel makes every new generation of HDD development more interesting, particularly from the viewpoint of deploying modern modulation and coding techniques.

The digital magnetic recording channel is inherently nonlinear because of the *M-H* hysteresis loop. That is, scaling the input current does not proportionately scale the output voltage since the remnant magnetization does not change appreciably for a large increase in the applied field. However, for a fixed write current that is sufficiently large to saturate the magnetic medium, the output signal can be constructed as a linear combination of the response due to the individual inputs symbols. Thus, in a limited regime of operation, the write process is nonlinear, but the readback process is linear. The output (readback) voltage waveform can be written as a pulse amplitude modulated (PAM) signal:

$$r(t) = \sum_k a_k h(t - kT) + \nu(t) \tag{4}$$

where $h(t)$ is the unit pulse response of the head/medium, $v(t)$ is the noise, and a_k is the sequence of input symbols forming the write current. Note that $a_k \in \{1, -1\}$. Using linearity once again, the unit pulse response can be written in terms of the more elementary response $s(t)$, called the transition response, as:

$$h(t) = s(t) - s(t - T) \tag{5}$$

where $1/T$ is the clock rate of the input sequence into the head/medium. The transition response corresponds to the head/medium response to a unit step change in

the write current polarity. Since the digital magnetic recording channel is peak amplitude limited, the peak of the transition response represents the maximum value of the output signal. The average power of the channel output for random data depends upon the operating density.

The readback signal can also be written in terms of the transition response as follows:

$$r(t) = \sum_k b_k s(t - kT) + \nu(t) \tag{6}$$

where $b_k = (a_k - a_{k-1})$ is the sequence of data transitions. Since a_k is binary, b_k is ternary ($b_k \in \{2, 0, -2\}$), where $b_k = -2$ denotes a change in write current polarity from positive to negative, $b_k = 0$ denotes no change, and $b_k = 2$ denotes a change from negative to positive. Note that successive nonzero data transitions alternate in polarity. Based on analytic results, the step response in digital magnetic recording channels is commonly modeled by a Lorentzian pulse given by

$$s_L(t) = \frac{A_L T}{\pi t_{50}} \frac{1}{1 + \left(\frac{2t}{t_{50}}\right)^2} = \frac{A_L}{\pi \delta} \frac{1}{1 + \left(\frac{2t}{\delta T}\right)^2} \tag{7}$$

where t_{50} is the width of the step response at half its maximum amplitude, A_L is the peak amplitude scaling factor, and δ is the *normalized linear density*, defined as:

$$\delta = \frac{t_{50}}{T} \tag{8}$$

The parameter t_{50} measures the temporal dispersion of the step response.[7] The model of Eq. (7) assumes that the head gap and the head/medium spacing are zero. More elaborate models are also available (see [5]), but the single-parameter Lorentzian pulse model is adequate for investigating the relative performance of different signal processing and coding schemes. The normalized linear density measures the number of bit cells that are packed per t_{50}, the half-amplitude-pulse-width, and is used as the parameter for comparing different detection methods. Today's HDD products have values of δ ranging from 2.3 to 3.0. Even with the application of zoned recording, the normalized linear density varies from the inner radius to the outer radius.

Figure 6 shows the Lorentzian transition response and Fig. 7 shows the corresponding pulse responses for varying values of δ. Both responses are symmetrical, and thus they have linear phase characteristics. The intersymbol interference (ISI) causes the amplitude and the energy of the pulse response to decrease as the linear density is increased.

In the frequency-domain, the amplitude spectra of the Lorentzian pulse and transition responses are given by:

$$H_L(\Omega) = j2TA_L \sin(\pi\Omega) \exp(-\pi\delta|\Omega|) \tag{9}$$

[7] When measured spatially, the spatial dispersion $pw_{50} = vt_{50}$ where v is the linear velocity of the disk. Many reported publications define the normalized linear density as pw_{50}/T.

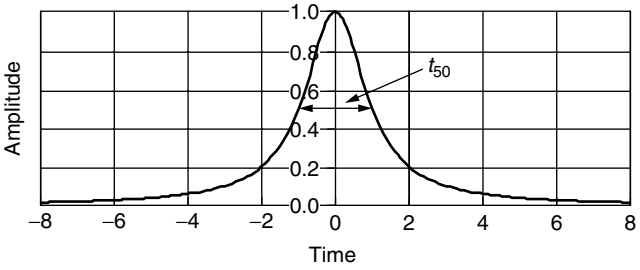


Figure 6. Lorentzian transition response.

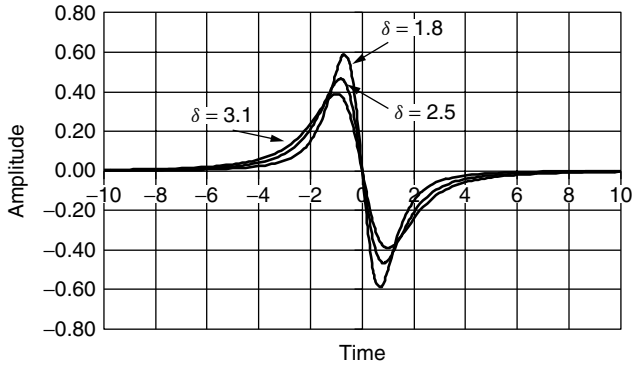


Figure 7. Lorentzian pulse response for varying linear density.

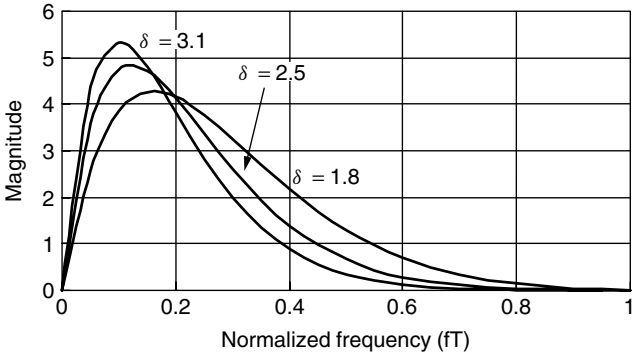


Figure 8. Amplitude spectra of lorentzian pulse.

and

$$S_L(\Omega) = A_L \exp(-\pi\delta|\Omega|) \tag{10}$$

where $\Omega = fT$ is the normalized frequency. Figure 8 shows the amplitude spectra of the Lorentzian pulse for different values of δ. The high-frequency content of the signal spectrum becomes increasingly attenuated because of the increased ISI at higher linear densities. Note that the amplitude spectrum extends beyond the Nyquist frequency ($f = 1/2T$), thereby requiring special consideration of sampling phase selection in symbol-spaced finite impulse response (FIR) equalizers. The phase spectrum of the pulse response is linear.

In summary, from the signal perspective, the digital magnetic recording channel is band-limited with a peak amplitude constraint, instead of the average power constraint generally associated with most communications

channels. Severe ISI occurs as the linear density is increased for a given recording channel, resulting in loss of pulse energy. The readback signal is corrupted by channel impairments, some of the major ones of which are outlined below. For a more detailed and exhaustive treatment, please refer to Refs. 1, 5.

In the digital magnetic recording channel, the *noise* sources include the following:

Media noise depends on the type of media. In particulate media, the noise is due to statistical distribution of the magnetic particles. It is modeled as additive, Gaussian, stationary, and with power spectrum similar to that of the signal. In thin film media, which are ubiquitous in today's disk drives, the noise is due to the randomness in the width of the recorded transitions. Figure 9 illustrates the source of this noise. As shown, the recorded transitions are far from being a straight line. Instead, they exhibit a zig-zag microstructure with a shape that varies randomly with each transition. The nominal transition response and its location then depend on the average width, w, and the average center of the recorded transitions. The statistical variation from these nominal values constitutes media noise. It is, in general, data-dependent and neither stationary nor additive. However, under some simplifying yet realistic assumptions, it can be modeled as additive and stationary (see Appendix 2C in [4]). *Head Noise* also depends on the head type. In MR heads, it is due primarily to the thermal resistances within the MR element and its contacts, and is, therefore, modeled as additive, white, and Gaussian. *Preamplifier noise* is added to the readback signal during its amplification. It is due to the electronic circuits in the signal path, and is also modeled as additive, stationary, Gaussian, and largely white.[8]

The above noise sources are mutually uncorrelated and their relative mix depends on the data rate and the magnetic and mechanical parameters of the head/media interface. For present-day systems, the media noise power is 1–4 dB higher than that of the electronics noise, which increases from the inner radii to the outer radii as the data rate is increased in zone-bit recording.

Inter-track interference (crosstalk) is caused by the pick up of signals from adjacent tracks as the head moves offtrack during the retrieval process. These interference signals are mitigated with accurate servo positioning, which attempts to keep the read head in the center of the track with very high probability. Further mitigation of interference signals is achieved by making the inductive write head wider than the MR read head. This so-called write-wide, read-narrow concept effectively introduces an additional guard band between tracks. The adjacent track interfering signal is, of course, a filtered version of the recorded data, and hence neither stationary nor Gaussian.

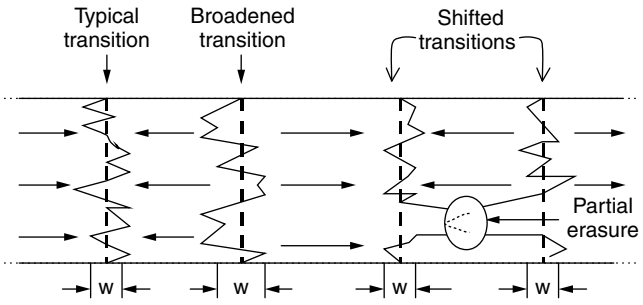


Figure 9. Recorded transitions in thin film media.

In addition to noise and interference, the readback signal may be corrupted by *distortion*, which may be linear or nonlinear, and generally grows as the recording density is increased. Some of the sources of nonlinear distortion are outlined below:

Nonlinear distortion in the form of *transition shift* occurs when the recorded transition is shifted from its intended location. Such shifts may occur when adjacent transitions get too close and bandwidth limitations in the write path, resulting in inadequate rise times of the write current or of the flux in the head gap, cause the transitions to move. Similarly, since successive transitions alternate in polarity, the magnetic field from the preceding transition[9] can interact with that of the new transition to aid its recording, thereby shifting the new transition earlier than intended. The amount of shift depends heavily on the operating conditions of the head/medium, and decreases rapidly as the minimum transition spacing increases. Such transition shift phenomenon is typically limited to transitions which are one symbol duration apart, but it can extend to two or more symbol durations if the linear density is very high. In practice, this nonlinear shift is virtually removed by: (1) ensuring adequate rise times in the write path, and (2) using precompensation during data recording wherein selective transitions in the write current are "delayed" by a prescribed amount to offset the subsequent "shift-early" effect.

The above transition shift is due to field interactions involving the new data sequence being recorded; similar shifts can occur because of residual fields from previous recordings. Because there is typically no dedicated erasure cycle in magnetic recording, the field from previously recorded data, especially those associated with low-frequency patterns, can "impede" or "aid" the recording of the new transition, causing it to shift. This effect is mitigated through careful design of the head/media parameters to achieve a prescribed "overwrite" signal-to-noise ratio, which guarantees a prescribed power ratio between the new readback signal and that from a previously recorded pattern.

Nonlinear distortion can occur with MR heads during readback. In single stripe MR head configurations,[10] which are widely deployed today, the stripe is biased to achieve a linear transfer characteristic between the change in flux and the associated change in MR resistance. Because of tolerances in MR stripe and the bias point, perfect linearity

[8] Some roll-off in the noise spectrum may occur at the upper and lower edges of the readback signal spectrum depending upon the amplifier design.

[9] This field is often referred to as demagnetizing field, which essentially has the effect of lowering the coercivity.

[10] Single stripe MR heads produce single-ended readback voltage signal; with dual stripe MR head, differential combining may be done to circumvent this effect.

is not achieved and some amount of memoryless, quadratic nonlinearity is introduced. The resulting signal has pulse asymmetry wherein the negative and the positive pulses may have different heights or widths, or both. This effect may be compensated by adaptively canceling the quadratic term before detection.

Nonlinear intersymbol interference can also occur in thin film media as recorded transitions get too close to each other. As noted earlier, the microstructure of the recorded transition in thin film media has a zig-zag signature (see Fig. 9). At very high linear density, portions of successive zig-zags may merge, causing the transitions to "weaken" and the readback signal amplitude to become smaller than that predicted by the linear model. This effect is referred to as *partial erasure*. It can be mitigated using precompensation during data recording wherein write current transitions separated by one symbol duration are moved away from each other by some prescribed amount.

In addition to the above nonlinear distortions, transient disturbances due to imperfections of the media may distort the readback signal. Media defects can cause "dropouts" in the readback signal amplitude. Such defects are screened during surface analysis of the media at the time of manufacturing of the disk drive. The defective sectors are precluded from storing information by the drive controller. Another form of distortion, referred to as *thermal asperity*, occurs when the single-stripe MR head bumps against a high spot on the medium. A large voltage transient is created because of the heating of the MR element, causing the small readback signal to modulate the transient signal. The transient decays exponentially as the MR element returns to its ambient temperature. This effect is mitigated during data retrieval by detecting the onset of the transient at the very earliest stage of the signal processing chain to avoid saturation of the subsequent blocks and loss of synchronization. Since the energy of the transient signal is located near the lower band-edge of the signal spectrum, the lower corner frequency of the receive filter is temporarily increased to filter out the ensuing transient. Such an approach is effective in limiting the span of the data errors to the correction capability of the error correcting code.

Channel identification based on the use of pseudo-random binary sequences has been developed to isolate the various nonlinear distortion effects outlined above. This method can be used in near real-time to define the precompensation parameters to linearize the channel (see [22]). As discussed in the next section, the signal processing methods deployed in magnetic recording assume the channel to be linear.

Linear distortion in the form of intersymbol interference (ISI) is by far the major contributor of distortion at high linear density, resulting in reduced energy and attenuated high-frequency content of the pulse response. The application of classical communication techniques to high density digital magnetic storage has been investigated over the past two decades, culminating in the development of many new coding and signal processing techniques that address the unique requirements of data storage (see [32,33]). Some of those techniques are discussed in the next section.

4. SIGNAL PROCESSING AND CODING METHODS

Signal processing and coding methods have played a vital role in digital magnetic recording systems, especially during the past decade as the bit cell dimensions have shrunk at an accelerated pace, requiring detection methods that are bandwidth and SNR efficient. Some of the methods used commercially are described in this section.

Figure 10 shows a block diagram of the data channel for digital magnetic recording. On the recording ("transmit") side, the data to be recorded are first encoded using an error-correction code (ECC), which is typically based on Reed-Solomon codes. The encoded output is applied to a modulation code and an associated precoder to achieve prescribed properties in the readback signal during data retrieval. The modulation code adds redundancy to improve signal detectability, but the precoder performs a one-to-one mapping on the encoded sequence. Different modulation codes have been used over the years along with different precoders, as discussed below. The encoded output is used to generate the write current waveform, which is then applied to the pre-compensation circuit to suitably time-shift the transitions associated with prescribed data patterns. As noted previously, symbol-spaced transitions are typically preshifted to linearize the recording channel.

On the retrieval side, the readback signal is amplified and then applied to the receiver, which is often referred to as "Read Channel" within the data storage community. The Read Channel is similar to a typical digital baseband communication receiver, comprising blocks that perform the functions of gain control, timing control, synchronization, receive filtering, equalization, detection, and decoding. In addition, compensation techniques may be incorporated to address other impairments, such as nonlinearity in MR heads and thermal asperity processing. This section will cover the evolution of the detection methods but not discuss important receiver functions like timing recovery, gain control, and synchronization. The reader is referred to Ref. 8 for a detailed treatment of those functions.

Figure 11 shows the evolution of the coding and detection methods in digital magnetic HDD. The upper legend in the box denotes the modulation code, and the lower legend denotes the detection method. The dashed box, denoting "Turbo Coded EPRML/GPRML," is not deployed commercially at the time of this writing. But, as

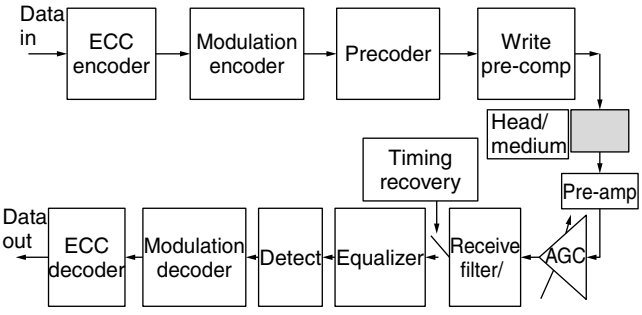


Figure 10. Magnetic data storage channel.

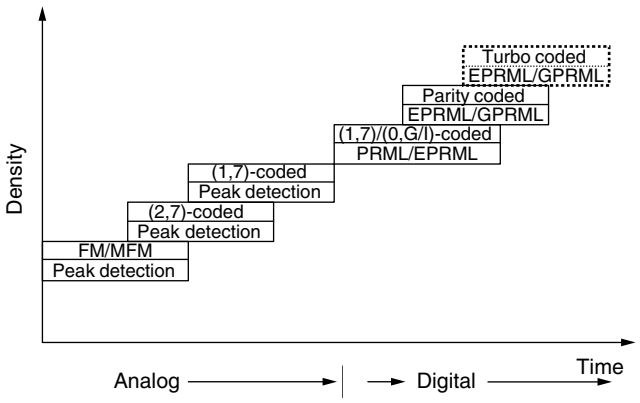


Figure 11. Evolution of coding and signal processing techniques.

with other communication channels, deployment of Turbo Coding in digital magnetic storage is of intense interest to researchers and practitioners alike, and simulation results (see [21]) to date show it to be highly effective in providing SNR benefits over other channel coding methods.

The Peak Detection method, relying on analog signal processing, was used ubiquitously in disk drives for over three decades to recover the recorded data from the analog readback signal corrupted by various impairments outlined in the previous section. It is based upon the simple observation that, in the absence of any inter symbol interference (ISI), the maximum (minimum) value of the transition response coincides with the location of the recorded transition, which, in turn, corresponds to a change in the polarity of the write current. Using this observation and the NRZI format[11] to represent the write current, the Peak Detection method detects the presence or absence of the signal peak within each symbol interval. A peak is considered to be present if the applied signal exceeds a prescribed threshold *and* its derivative has a zero crossing. Otherwise, the peak is considered to be absent. Thus, the presence of a peak denotes a "1" and the absence denotes a "0."

The performance of Peak Detection degrades rapidly in the presence of ISI because: (1) the signal peaks shift away from the location of the transitions, and (2) the peak amplitude associated with closest transitions decreases (see Fig. 7). The effect of ISI is mitigated with a class of modulation (line) codes called run-length limited (RLL) *codes*. These codes prescribe run-length constraints on recorded sequences to extend the applicability of Peak Detection. The run-length constraints are typically designated as (d, k), where d and k are nonnegative integers (with $d < k$) that denote, respectively, the minimum and the maximum number of "0s" between "1s" at the encoder output. For example, the (1,7) RLL code produces sequences that contain at least one "0" and at most seven "0s" between any pair of "1s." When combined with the NRZI format, wherein "1" produces a polarity change in the write current, these parameters determine the minimum $(= (d + 1))$ and the maximum $(= (k + 1))$ symbol intervals between recorded transitions. Since transitions produce signals with nonzero amplitudes, the k constraint acts to ensure that corrective updates are available at some minimal rate for the timing recovery and the automatic gain control loops. Similarly, the d constraint controls the separation between closest transitions, and thus the resulting ISI, if any. Together, the (d, k) pair defines the highest code rate, called the code capacity, which can be achieved for the prescribed constraints. The d constraint can be removed by setting $d = 0$; likewise, the k constraint can be removed by setting $k = \infty$. For a detailed description of RLL codes and their construction, refer to Ref. 19.

Early HDD systems were based on rate 1/2 codes, with the run-length constraints evolving from (0,1) for frequency modulation (FM) to (1,3) for modified-FM (MFM), to (2,7). By progressively increasing d, these codes achieved higher linear densities with peak detection without incurring a code rate penalty or performance degradation. The approach was, however, not extendible to $d = 3$, since such a constraint could not be achieved with a rate 1/2 code.[12] Instead, the (1,7) code with rate 2/3 was adopted. The ISI increased because of the $d = 1$ constraint, but the symbol duration also increased because of the higher code rate, resulting in more available energy for distinguishing signals most likely to be confused. With suitable equalization to mitigate the ISI, the (1,7) code provided a net performance gain over its predecessor (2,7) code. The equalization was based on the simple approach of boosting the high frequencies in the readback signal to slim the transition response.

As the magnetic bit cell continued to shrink, the combination of equalization, RLL coding, and peak detection was no longer adequate to achieve acceptable performance; new detection methods were required to cope with decreasing SNR and severe ISI. Classical transmission techniques, including partial response signaling [4,8,9,25,26,29], decision feedback equalization and its variations [2,3], along with powerful modulation coding [16] were investigated. An exhaustive treatment of the many results (see Refs. 32, 33) from these investigations is beyond the scope of this chapter. Today, partial response signaling is deployed in HDD systems ubiquitously. The remainder of this section is devoted to the theory and practice of partial response signaling in digital magnetic recording channels.

The partial response signaling concept as applied in the digital magnetic recording channel is illustrated in Fig. 12. The readback signal is suitably equalized to a target partial response signal and sampled before detection. Since the input to the write block is a sequence of data symbols, the entire signal path, comprising the write/read/pre-amplify/equalizer/sampling blocks, can be represented by a discrete-time transfer function based on the choice of the target partial response signal. The

[11] The non-return-to-zero-invert (NRZI) format inverts the write current polarity with every occurrence of "1," thereby causing a transition to be recorded on the medium. It is a form of precoding commonly referred to as differential encoding in data communications.

[12] The code capacity for $d = 3$ and unconstrained-k code (that is, the $(3, \infty)$ code) is 0.4650. It is even lower for a code with a finite k constraint.

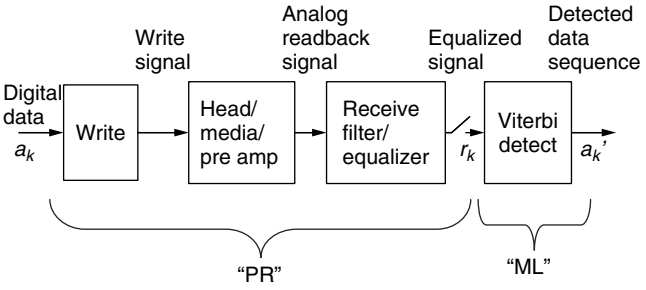


Figure 12. The PRML concept.

detector is based on the Viterbi algorithm performing maximum-likelihood sequence estimation. The overall approach is referred to as partial response maximum-likelihood (PRML) in the HDD industry, where the "PR" and "ML" parts are identified in the figure. The relevant partial response targets for the digital magnetic recording channel are discussed below.

Partial response signaling was developed for transmission of data over band-limited channels (see Ref. 18). Using the sampling theorem, any partial response signal u(t) can be expressed as:

$$u(t) = \sum_k u_k \frac{\sin[\pi(t - kT)/T]}{[\pi(t - kT)/T]} = \sum_k u_k \text{sinc}[(t - kT)/T] \tag{11}$$

where u_k represents the sample value u(kT) and the function sinc(y) is defined as the ratio $\sin(\pi y)/\pi y$. In the frequency domain, the spectrum of the partial response signal is given by:

$$U(f) = \sum_k u_k \exp(-j2\pi fkT), \quad |f| \le 1/2T \tag{12}$$

Partial response signals are designed to support a symbol rate of $1/T$ over a bandwidth of $1/2T$ Hz. With the binary input constraint for the digital magnetic recording channel, this represents a spectral efficiency of 2 bits/sec/Hz. By defining the transform of the unit delay operation, $D = \exp(-j2\pi fT)$, Eq. (12) can be written as a polynomial in D, given by[13]

$$U(D) = \sum_k u_k D^k \tag{13}$$

Infinitely many pulse shapes can be created by choosing different values of u_k in Eq. (11). In general, the number of nonzero u_k's is minimized to achieve the desired performance objectives at least cost.

To understand which partial response signals are well suited for digital magnetic recording, consider the model of saturation recording again. The differencing operation inherent in the recording process (see Eq. (5)) suggests that $(1 - D)$ must be a factor in the polynomial defining the target pulse response for the channel. The $(1 - D)$ partial response system has a high-pass amplitude spectrum with a null at dc. The low-pass filtering effect during readback, due to the gap between the head and the medium, can be modeled by the $(1 + D)^n$ partial response signals, where n is a nonnegative integer. The combined polynomials, representing a set of bandpass responses, are given by

$$P_n(D) = (1 - D)(1 + D)^n \tag{14}$$

These polynomials represent a class of partial response targets that are well suited for the digital magnetic recording channel. This class is called extended partial response (EPR) systems in the magnetic recording literature, where $n = 1$ is referred to as PRML, $n = 2$ as EPRML (for Extended-PRML), $n = 3$ as E^2PRML, and so on. The polynomial $(1 - D^2)$ corresponding to $n = 1$ is the well-known Class IV or Modified Duobinary partial response system [18]. Its application to digital magnetic recording was first noted in [17].

Note that, while $P_n(D)$ defines the target pulse response, the polynomial $(1 + D)^n$ can be interpreted to represent the target transition response since the $(1 - D)$ factor models the differencing operation in Eq. (5). The sample values of the target transition response are given by the binomial coefficients:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{15}$$

since

$$(1 + D)^n = 1 + \binom{n}{1} D + \binom{n}{2} D^2 + \cdots + \binom{n}{n} D^n \tag{16}$$

Note that for large n, the transition response is approximately Gaussian. Indeed, MR heads typically exhibit a transition response between a Lorentzian pulse and a Gaussian pulse. The sample values of the pulse response, p_k, can be derived from the binomial coefficients. Figure 13 shows the target EPR pulse shapes for $n = 1, 2, 3$, and 4. Qualitatively, the EPR signals with increasing n are similar to the Lorentzian signals with increasing δ (see Fig. 7).

Referring back to Fig. 12, $P_n(D)$ defines the "PR" part of the system; that is, it defines the input-output relationship of the sampled data sequences. Thus, if $\{a_k\}$ represents the input data sequence, the noise-free output sequence $\{y_k\}$ is given by:

$$Y(D) = P_n(D)A(D) \tag{17}$$

where

$$Y(D) = \sum_k y_k D^k \tag{18}$$

$$A(D) = \sum_k a_k D^k \tag{19}$$

and

$$P_n(D) = \sum_{k=0}^{n+1} p_k D^k \tag{20}$$

[13] The D is replaced by z^{-1} in digital signal processing literature. The two are equivalent transform representations of a sample sequence.

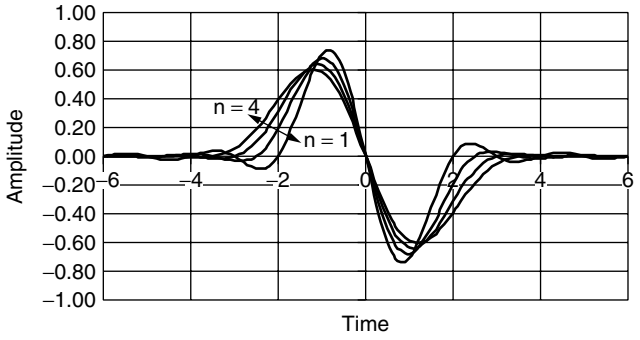


Figure 13. Target EPR pulse shapes for $n = 1, 2, 3$, and 4.

In the time-domain, the noise-free target sampled output can be written as:

$$y_k = \sum_{i=0}^{n+1} p_i a_{k-i} = p_0 a_k + p_1 a_{k-1} + \cdots + p_{n+1} a_{k-n-1} \qquad (21)$$

where the set $\{p_i\}$ is derived from the binomial coefficients. The above formulation stipulates a finite impulse response model for the equalized magnetic recording channel where controlled ISI is allowed between the responses due to the current and the $(n + 1)$ previous inputs. The controlled ISI is prescribed by the choice of $\{p_i\}$, the sample values of the target pulse response. Because of the controlled ISI, the number of output levels is greater than the number of input levels, and depends on the target partial response polynomial. The sampled input to the detector is the noisy sample, given by:

$$r_k = y_k + v_k = \sum_{i=0}^{n+1} p_i a_{k-i} + v_k \qquad (22)$$

where v_k is the sampled noise. The signal spectrum, $S_n(f)$, for each $P_n(D)$ is obtained by setting $D = \exp(-j2\pi fT)$ in Eq. (14), yielding

$$S_n(f) = jT2^n \cos^{n-1}(\pi fT) \sin(2\pi fT), \quad |f| \le 1/2T \qquad (23)$$

Figure 14 shows a plot of the amplitude spectrum $|S_n(f)|$ for different n. Since the factor $(1 + D)$ has the effect of introducing a null at the Nyquist frequency, higher values of n introduce higher order nulls, thereby attenuating the high-frequency content in the target response. This behavior is similar to that exhibited by the Lorentzian model shown in Fig. 8 as δ is increased. As discussed below, for a given linear density, the order of the EPR polynomial can be chosen to maximize the available SNR at the detector.

The equalization of the readback signal to the EPR target can be implemented using analog or digital filters, or combinations thereof. Even with zone bit recording, the linear density changes radially, thus requiring some level of adaptation, either real-time or during zone switching, of the equalizer response. Commercially, both analog and digital implementations have been deployed successfully. The optimization of the filter parameters is generally based on the minimum mean-squared error (MMSE) criterion (see Ref. 24).

Analog equalization typically is implemented using a continuous-time filter with linear phase response. The filter comprises a cascade of two real-axis zeros and a low-pass filter, typically the 7th order Bessel filter. The location of the zeros and the cutoff frequency of the low-pass filter are jointly optimized for each zone and preset by the HDD controller during zone switching. This approach was deployed commercially with EPR and E^2PR target signals (see Refs. 7, 23).

Discrete-time equalization is implemented with a programmable or adaptive finite impulse response (FIR) filter. The choice of the sampling phase of the unequalized readback signal is important to the error rate performance when using symbol-spaced FIR filters. As noted in Fig. 8, the readback signal spectrum typically extends beyond the Nyquist frequency. When sampled at the symbol rate before equalization, foldover of the readback signal spectrum occurs about the Nyquist frequency, resulting in the well-known aliasing effect. The folded spectrum determines the noise enhancement penalty depending on whether the aliasing is additive or subtractive, which, in turn, depends upon the sampling phase. Figure 15 illustrates this effect for the Lorentzian channel with $\delta = 2$. Phase (1) corresponds to sampling at the peak of the transition response, resulting in additive aliasing and no null at Nyquist frequency. Phase (2) corresponds to sampling at $\pm T/2$ seconds from Phase (1), resulting in a null at Nyquist frequency. Even though the target EPR signals require a null at Nyquist frequency, the folded spectrum without the null yields better error rate

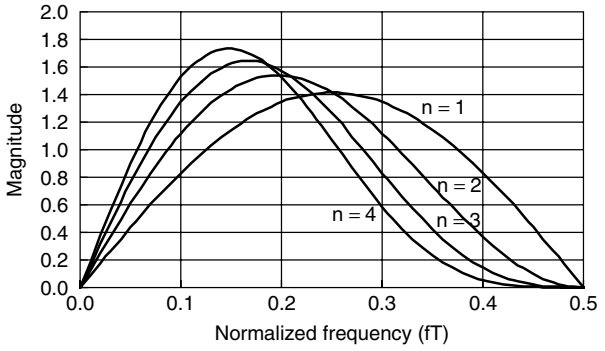


Figure 14. Amplitude spectra of the EPR pulse response signals.

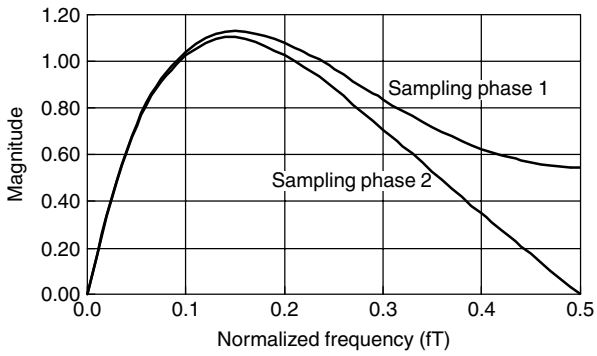


Figure 15. Folded Lorentzian spectrum for two different sampling phases.

performance because of less noise enhancement in the equalizer. Theoretical and experimental results show that the choice of sampling phase can, depending on the operating linear density, affect the SNR by 1–2.5 dB at the detector for PRML and EPRML (see Ref. 27). For higher order systems, the effect is less since the pulse energy of the target response is concentrated at the lower frequencies. The optimum sampling phase depends on the phase response of the channel, including the receive filter. With a linear phase receive filter and symmetric transition response, the optimal sampling phase typically corresponds to sampling at the peak of the transition response.

The controlled ISI inherent in partial response signaling introduces structure in the equalized waveform that can be used to unravel the ISI with little or no loss in performance relative to ISI-free signaling. The required optimum detector is based on maximum-likelihood (ML) sequence estimation instead of symbol-by-symbol detection, as is the case with peak detection. The ML estimation relies on the Viterbi algorithm, which takes the noisy samples r_k in Eq. (22), output from the equalizer, and determines the most likely recorded sequence $\{a_k'\}$. The estimation is based on minimizing recursively the squared-error between the r_k sequence and all *allowed* y_k sequences from the beginning to the end of the data sector. The recursive procedure is applied to a trellis diagram, which graphically depicts all possible states of the channel, defined by the $(n+1)$ most recent inputs $\{a_{k-1}, \ldots, a_{k-n-1}\}$, and the allowed transitions between channel states from time k to $(k+1)$. The complexity of the Viterbi algorithm is proportional to the number of states, which, for the EPR signals, equals 2^{n+1}. For details of the Viterbi algorithm, refer to Ref. 24.

Modulation (line) coding with run-length constraints is also needed with partial response signaling. However, unlike peak detection, the d constraint can be zero since partial response signaling is fundamentally based on allowing controlled ISI at the detector input. The purpose of modulation coding then becomes: (1) to provide frequent updates to the timing recovery and the automatic gain control loops, and (2) to facilitate survivor path merges within a prescribed length of the path memory in the Viterbi detector. The first requirement is similar to that encountered in all digital communication receivers. The second requirement stems from the observation that the EPR polynomials have nulls at both dc and Nyquist frequency. Thus, data sequences with contiguous 1s or -1s (dc), or alternating 1s and -1s (Nyquist frequency) produce zero output levels, and are hence indistinguishable. Such sequences traverse paths in the Viterbi trellis that do not merge, nor accumulate the minimum Euclidean distance, which determines the performance of the ML sequence detector (see Eq. (24)). To avoid performance degradation due to unmerged survivor sequences, the maximum run-length of like symbols (1s or -1s) in both global (contiguous) and interleaved strings of recorded data are constrained to at most G and I symbols, respectively γ. The resulting modulation code is designated (0, G/I), where the 0 represents the d constraint. The G and I constraints may be achieved by interleaving two $(0, k)$ RLL codes. However, a tighter G constraint is achieved for a given code rate by designing the (0, G/I) code as a single code. The (0, G/I) modulation codes typically use the interleaved NRZI (I-NRZI) format to represent the write current. The I-NRZI precoder is based upon applying NRZI precoding to the even and odd bits of the encoded data sequence independently. The first PRML Read Channel deployed commercially in HDD was based upon a rate 8/9, (0, 4/4) code. Subsequently, rate 16/17 codes with looser G and I constraints were used with PRML and EPRML systems.

To ascertain the relative performance of the EPR systems, consider first a channel that already has the prescribed EPR response without any equalization. With additive white Gaussian noise as the only impairment, the probability of error at moderate to high SNR with random data input is given by (see Ref. 12):

$$P_e = K_1 Q\left(\frac{d_{\min}}{2\sigma}\right) \tag{24}$$

where $d_{\min}$ is the minimum Euclidean distance for an error event in the Viterbi detector, K_1 is the average number of such error events, and Q is the area under the tail of the Gaussian density function, given by:

$$Q(z) = \frac{1}{\sqrt{2\pi}} \int_z^\infty \exp(-q^2/2)\, dq \tag{25}$$

If the given EPR channel is used to transmit just one pulse, then no ISI is present. The probability of error for such single use of the channel is given by

$$P_{MF} = K_2 Q\left(\frac{|p|}{2\sigma}\right) \tag{26}$$

where

$$|p|^2 = \sum_{k=0}^{n+1} {p_k}^2 \tag{27}$$

is the energy of the pulse response and K_2 is the number of ways the pulse is incorrectly detected. P_{MF} denotes the probability of error in detecting the isolated pulse (without ISI) with the optimum linear receiver, namely, the matched filter receiver. Comparing Eqs. (24) and (26), the ratio ${d_{\min}}^2/|p|^2$ represents the loss due to the controlled ISI. It is the fraction of the pulse energy that the Viterbi detector is able to use toward discriminating the sequences most likely to be confused. The ISI loss and some key parameters for EPR systems are listed in Table 1. Note that the $n = 1$ and $n = 2$ systems do not incur any ISI loss; thus, they perform as well as an ISI-free signaling scheme. The ISI loss column also represents the relative SNR loss in the detector in choosing a higher order EPR polynomial. From this perspective, $n = 1$ and 2 require the same SNR at the detector for a desired error rate, whereas $n = 3$ requires 2.2 dB more, and so on. The number of output levels for binary inputs and the t_{50}/T of the target transition response, representing the best linear density match in white Gaussian noise, are also given in Table 1.

In deploying EPR systems, complexity and performance considerations alone dictate that the degree of the polynomial, determined by the choice of n, be as small as

Table 1. Key Parameters of EPR Systems

n	No. of Output Levels	$\|p\|^2$	$d^2_{\min}$	$d^2_{\min}/\|p\|^2$	ISI Loss (dB)	t_{50}/T
1	3	2	2	1	0	1.6
2	5	4	4	1	0	2.0
3	7	10	6	0.6	2.2	2.3
4	13	28	12	0.4	3.7	2.6

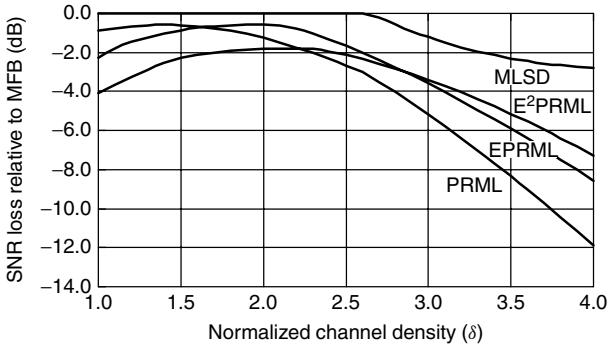


Figure 16. Performance of EPR systems on Lorentzian channel.

possible. Indeed, higher order partial response polynomials have not been of much interest in communication channels. But the digital magnetic recording channel is different in that the transition response is fixed for a given head/medium combination. As the linear density is increased with a given EPR target, the noise enhancement penalty from the equalizer increases because of the growing mismatch between the readback signal and target signal spectra. Indeed, when the ISI loss and the noise enhancement penalty are taken into account, a given polynomial will provide acceptable performance over a range of recording density; beyond that range, a higher degree polynomial is needed to reduce the noise enhancement penalty, and achieve acceptable performance. Figure 16 illustrates this point using the Lorentzian pulse generator followed by additive white Gaussian noise as the model for the recording channel. Using the MMSE criterion for the equalizer design and taking into account the effect of noise correlation in the Viterbi detector, the asymptotic SNR loss relative to the matched filter bound ($\mathrm{SNR}_{\mathrm{MF}} = |p|^2/\sigma^2$) is plotted as a function of normalized density δ. Note that 0 dB loss represents matched filter performance. As shown, PRML ($n = 1$) provides better performance for δ in the range of 1.0 to 1.6, EPRML ($n = 2$) provides better performance for δ in the range of 1.6 to 2.8; and so on. To provide a benchmark, Fig. 16 also shows the performance of the maximum likelihood sequence detector (MLSD), which is the optimum detector for ISI channels (see Ref. 12). As shown, the MLSD detector achieves the matched filter bound up to the normalized density of approximately 2.6; that is, its performance is the same as that of an ISI-free channel. Beyond that range, however, the performance of MLSD degrades relative to the matched filter bound because of the increasing ISI. Note that the performance of the EPR systems, featuring moderate complexity, is within 1–2 dB of MLSD over the range of linear densities of current interest.

Read Channel products based on PRML, EPRML, and *E²PRML* have been deployed commercially in hundreds of millions of magnetic disk drives during the past decade. More recently, there has been increased interest in the development of *generalized partial response* (GPRML) polynomials to bridge the performance gap between MLSD and EPR polynomials. Such polynomials are derived using search procedures that minimize the probability of error on an empirical model of the recording channel. Since the optimum target response is one which whitens the noise at the detector input, the GPRML schemes, in a sense, perform spectral matching and noise whitening jointly with a polynomial of prescribed degree and spectral null constraints.[14] Unlike the EPR targets, which are symmetrical like the ideal channel, the resulting GPR targets are asymmetrical and closer to a minimum phase representation of the channel.[15] The GPRML approach is shown to be quite promising at high linear density ($\delta > 2.7$) with polynomials of degree 4 and above. About 1 dB SNR advantage is achieved over an equivalent EPR polynomial with modest increase in the equalizer complexity (see Ref. 10).

As recording densities continue to increase, *channel coding* techniques have been developed for partial response signaling to deal with the reduced SNR. Unlike communication channels, however, channel coding is difficult to apply in the digital magnetic recording channel because of the binary constraint on the input waveform. Also, for a given head/medium combination, the pulse-energy-to-noise ratio, representing the matched filter bound, decreases rapidly because of the increased ISI and noise bandwidth associated with adding redundancy inherent in channel coding. The rate of this decrease is 6–9 dB per doubling of the symbol rate (or linear density). Thus, in order to achieve a net gain with a rate 1/2 code, the coding gain must be larger than 6–9 dB — a nontrivial task! Channel coding, therefore, must rely on the use of high code rates to eke out a net performance gain for a given recording channel.

Two types of channel coding methods have been developed and deployed commercially in HDD systems: *trellis coding* and *parity coding*. Both methods rely on suitably dealing with the most likely minimum distance error events for the target partial response signal, since it is these events that limit the performance of the detector (see Eq. (23)).

The underlying approach in the trellis coding method is similar to that used in data communications: increase the minimum Euclidean distance between all allowed sequences of output symbols, y_k. This objective is achieved by eliminating input data patterns that support the prescribed minimum distance error events for a given partial response system. The constraints on the input sequence together with those from the target partial response signal are suitably combined to create a new trellis diagram which has larger minimum Euclidean distance than the uncoded system. Interestingly, the

[14] First order nulls at dc and at Nyquist frequency are typically retained, although some reported polynomials have dropped the null at the Nyquist frequency.

Table 2. Some Input Error Sequences for E^2PR

d^2	Input Error Sequence ($\pm e_k^a$)
6	1-11
8	1-11001-11 1-11-11-1 1-11-11-11-1... 1-11-11-11 1-11-11-11-11
10	1 1-110-11-1 1-11001-11001-11 ...

resulting trellis may, at times, be simpler than that for the uncoded system.

For example, Table 2 lists a few of the minimum distance input error sequences for the E^2PR polynomial. The *input error sequence* $\{e_k{}^a\}$ is the difference between any two possible input data sequences, and the *error event* $\{e_k{}^y\}$ is the difference between the corresponding noise-free channel output sequences (see Eq. (21)). The two error sequences are related as follows:

$$e_k{}^y = y_k{}^1 - y_k{}^2 = \sum_{i=0}^{n+1} p_i a_{k-i}{}^1 - \sum_{i=0}^{n+1} p_i a_{k-i}{}^2$$
$$= \sum_{i=0}^{n+1} p_i (a_{k-i}{}^1 - a_{k-i}{}^2) = \sum_{i=0}^{n+1} p_i e_{k-i}{}^a \qquad (28)$$

where $y_k{}^1$ and $y_k{}^2$ are two noise-free output sequences due to input data sequences $a_k{}^1$ and $a_k{}^2$, respectively. Since $a_k{}^i$ is binary, the associated error sequence alphabet $e_k{}^a$ is ternary, as given in Table 2. By avoiding the input NRZ sequences of the form 1-11 and −11-1, both the squared-distance 6 and 8 error events can be eliminated, along with some of the squared-distance 10 events. This constraint on the input sequence is easily introduced through a suitable combination of the $d = 1$ RLL code and precoding, and eliminating the associated states in the Viterbi trellis. The resulting trellis diagram has 12 states, instead of 16 states in the uncoded $E^2PRML(n = 3)$ system, along with minimum squared-Euclidean distance for an error event of 10. The resulting coding gain is $10\log(10/6) = 2.2$ dB. This combined coding and equalization scheme was deployed commercially to replace the (1,7)-Coded Peak Detection scheme.

Even higher code rates can be used on the E^2PR channel to achieve the coding gain of 2.2 dB. These codes rely on the use of a 16-state time-varying trellis to represent the constraints on the input sequences, and are referred to as TMTR (Time-Varying Maximum Transition Run-Length) codes (see Refs. 6, 20, 23). A more general and systematic trellis code construction method for partial response channels relies on suitably matching the nulls of the code spectrum with those of the partial response channel to increase the minimum Euclidean distance. The theory and implementation of this method can be found in Refs. 11, 16.

The parity coding method aims to detect and selectively correct the dominant error events at the output of the Viterbi detector to improve the error rate performance. Parity bits are suitably inserted within codewords to detect the occurrence of parity violations at the output of the detector. Soft decision correction, in the maximum likelihood sense, is applied to correct the parity violations. The applicability of this approach relies on the unique patterns that characterize the most likely error events for EPR and GPR targets. Unlike trellis coding, parity coding does not add any constraints on the input sequences beyond those imposed by the modulation code. The resulting code rate can be, therefore, higher than that for trellis codes. Indeed, in commercial deployments, the parity-coded schemes achieved the same code rate of 16/17 as in modulation-code-only systems by combining the parity and the modulation constraints in longer block codes,[16] such as rate 32/34 or rate 96/102.

To illustrate the parity coding approach, consider the use of EPRML on the Lorentzian channel with additive white Gaussian noise. Table 3 lists the ordered sets of error sequences as a function of the normalized linear density. The ordering is based on the likelihood of occurrence of each event, defined in terms of the SNR level above the most likely error event (normalized to 0 dB). The ordering of error sequences changes as a function of the linear density because of the changing noise correlation introduced by the equalizer. Note that, except for 1-11-1, all other error sequences involve an odd number of input bits. Thus, by appending a bit to create codewords with even parity, the occurrence of most likely error events within a codeword can be detected. The same approach can be applied to measured signal and noise from the recording channel and to other EPR or GPR targets. Indeed, the approach can be extended to include multiple bits of parity, wherein the data to be recorded is organized into multidimensional arrays and parity bits are suitably appended in each dimension to achieve a prescribed detection and correction capability.

Table 3. Input Error Sequences for EprmL System

Normalized Linear Density	Input Error Event Sequence	SNR Level (dB)
2.25	1 1-11-11 1-11-11-11 ...	0 0.19 0.33
2.5	1-11 1-11-11 1-11-11-11 ...	0 0.14 0.14
2.75	1-11 1-11-11 1-11-1 ...	0 0.45 0.58

[16] The resulting G and I constraints are looser relative to the modulation code with rate 16/17.

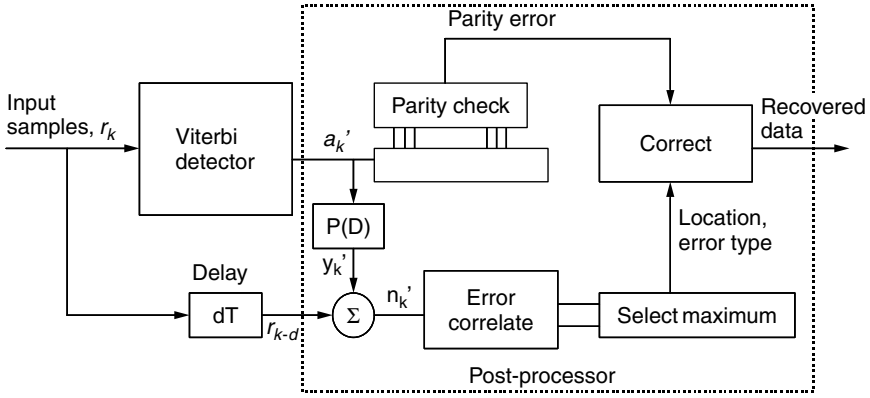


Figure 17. Detector structure for parity coding using postprocessor.

The detection and correction of parity violations may be performed within the Viterbi detector, or with a combination of the Viterbi detector and a postprocessor. With the former approach, the combined constraints of the parity code and the target response are incorporated into a time-varying trellis with twice the number of states of the original target response. The Viterbi algorithm is applied to the new trellis to recover the corrected data sequence. For a target response with 16 states, which represents the state of the art at the time of this writing, the increased complexity with this approach is quite significant. Instead, the combination of the Viterbi detector and postprocessor is commonly used with parity coding (see Refs. 10, 12, 31).

Figure 17 shows the postprocessor-based detector structure. The detection of parity violations is performed at the end of each codeword by the parity check block. The most likely location of the parity violation, if it occurs, is determined as follows: The Viterbi detector for the target partial response signal produces the estimated data sequence $\{a_k'\}$, which is used to reconstruct the noise-free partial response output symbols $\{y_k'\}$. Together with suitably delayed input samples, the sequence $\{y_k'\}$ is used to estimate the noise sequence $\{n_k'\}$, where $n_k' = r_k - y_k'$. These noise estimates are correlated at each bit time using a bank of matched filters, where each filter is matched to the most likely error event. The noise correlation outputs are only considered valid if the estimated sequence $\{a_k'\}$ supports the prescribed error events. The bit interval with the maximum valid noise correlation output is assumed to be the location of the error sequence. This location and the associated error event are then used to correct a parity violation, if necessary.

The parity coding scheme is effective in providing coding gains in excess of 1 dB without incurring any code rate penalty relative to systems with only modulation coding.

Unencumbered by signaling standards, the application of partial response signaling and channel coding techniques described above has been fast paced during the past decade. The peak detection method, which enjoyed widespread use for almost 30 years, was replaced by the more powerful PRML method in the early 1990s. PRML was completely displaced by EPRML and E^2PRML during the mid- to late 1990s. Parity and trellis coding techniques were introduced in late 1990s along with GPRML. Much research activity today is focused on combining Turbo coding techniques with partial response equalization (see Ref. 21). But issues related to the decoding delay inherent in Turbo decoding need to be resolved first. These issues pertain to delay requirements that exist within the datapath of the HDD as well as between the host CPU and the disk drive. Given the continuing thrust to double the areal density annually, it is just a matter of time before these issues are resolved. The trend in the semiconductor industry to integrate more and more functionality on a single chip is likely to also facilitate the development and deployment of Turbo coding methods.

5. CONCLUDING REMARKS

Digital magnetic storage has played a pivotal role in the evolution of storage systems over the past 45 years. Thousands of terabytes of storage are consumed annually worldwide, and the demand is exploding as new applications emerge in computing, communications, and entertainment systems. The underlying technologies in digital magnetic recording continue to progress at an exponential rate. And while some fundamental limitations due to the decreasing bit cell are anticipated above 100 Gbits/sq. inch, researchers are busy exploring means to overcoming those limitations. As noted in Ref. 28, no alternative storage technologies exist on the horizon which show promise for replacing the magnetic hard disk drive in the next ten years.

Over the past decade, equalization and coding methods have played a vital role in the growth of storage capacity per disk surface. New and powerful methods of combining equalization and coding are being developed and deployed commercially. This trend will continue as the digital magnetic channel itself continues to evolve based on new head, media, and recording technologies.

Acknowledgments

The author would like to convey special thanks to some colleagues from former DataPath Systems who, for six years, helped shape the trends in Read Channel technology and products beyond PRML: S. Altekar, J. Chern, C. Conroy, R. Contreras, L. Fang, Y. Hsieh, E. MacDonald, T. Pan, S. Shih, and A. Yeung; and to

former colleagues with whom he had the privilege of collaborating on modern signal processing and coding methods for digital magnetic storage: J. Cioffi, T. Howell, R. Karabed, M. Melas, A. Patel, P. Siegel, J. Wolf, and R. Wood. Thanks are also due to Ed Growchoski of IBM Corporation for providing the chart on the areal density trends.

BIOGRAPHY

Hemant K. Thapar received the M.S and Ph.D. degrees in Electrical Engineering from Purdue University in 1977 and 1979, respectively. He worked at Bell Telephone Laboratories, Holmdel (1979–84), at IBM Corporation, San Jose (1984–94), and at DataPath Systems (1994–2000), which he cofounded. He is presently a senior vice-president at LSI Logic Corporation, Milpitas, California, and an adjunct lecturer at Santa Clara University, California. He was corecipient of: the Best Paper Award at Interface 1984 for his work on high-speed, full-duplex data transmission; the Best Technical Report citation from IBM Almaden Research Center in 1989 for his work on PRML technology; and the 1991 IEEE Communications Magazine Prize Paper award for his paper on future technology directions in signal processing for data storage. Dr. Thapar is a Fellow of the IEEE and holds many patents and publications in the areas of digital communications, data storage, and networking. His technical interests are in the areas of data transmission and storage, networking, and VLSI architectures and design.

BIBLIOGRAPHY

1. T. C. Arnoldussen and L. L. Nunnelley, *Noise in Digital Magnetic Recording*, World Scientific Publishing Co. Pte. Ltd., Singapore, 1992.
2. P. S. Bednarz et al., Performance evaluation of an adaptive RAM-DFE read channel, *IEEE Trans. Magn.* **MAG-31**(2): 1121–1127 (March 1995).
3. J. W. M. Bergmans, Decisions feedback equalization for run-length limited modulation codes with $d = 1$, *IEEE Trans. Magn.* (1996).
4. J. W. M. Bergmans, *Digital Baseband Transmission and Recording*, Kluwer Academic, 1996.
5. H. N. Bertram, *Theory of Magnetic Recording*, Cambridge University Press, United Kingdom, 1994.
6. W. Bliss, An 8/9 rate time-varying trellis code for high density magnetic recording, *IEEE Trans. Magn.* **33**: 2746–2748 (Sept. 1997).
7. J. Chem et al., An EPRML digital read/write channel IC, *ISSCC Digest of Tech. Papers*, Paper 19.4, 320–322 (Feb. 1997).
8. R. D. Cideciyan, F. Dolivo, R. Hermann, W. Hirt, and W. Schott, A PRML system for digital magnetic recording, *IEEE J. Select. Areas Commun.* **SAC-10**(1): 38–56 (Jan. 1992).
9. J. M. Cioffi, W. L. Abbott, H. K. Thapar, C. M. Melas, and K. D. Fisher, Adaptive equalization in magnetic-disk storage channels, *IEEE Comm. Mag.* 14–29 (Feb. 1990).
10. T. Conway, A new target response with parity coding for high density magnetic recording channels, *IEEE Trans. Magn.* **34**(4): 2382–2386 (July 1998).
11. L. Fredrickson et al., Improved trellis coding for partial response channels, *IEEE Trans. Magn.* **31**: 1141–1148 (March 1995).
12. G. D. Forney, Jr., Maximum likelihood estimation of digital sequences in the presence of intersymbol interference, *IEEE Trans. Inform. Theory* **IT-18**(3): 363–378 (May 1972).
13. E. Grochowski and R. Hoyt, Future trends in hard disk drives, *IEEE Trans. Magn.* **32**(3): 1850–1854 (May 1996).
14. E. Grochowski, website: *http://www.storage.ibm.com/*
15. J. M. Harker, D. W. Bede, R. E. Pattison, G. R. Santana, and L. G. Taft, A quarter century of disk file innovation, *IBM J. Res. Devel.* **25**(5): 677–689 (Sept. 1981).
16. R. Karabed and P. Siegel, Matched spectral-null codes for partial response channels, *IEEE Trans. Inform. Theory* **37**(3): 818–855 (May 1991).
17. H. Kobayashi and D. T. Tang, Application of partial response channel coding to magnetic recording systems, *IBM J. Res. Devel.* **15**: (July 1970).
18. A. Lender, Correlative digital communication techniques, *IEEE Trans. Comm. Tech.* **COM-12**: (Dec. 1964).
19. B. Marcus, P. Siegel, and J. Wolf, Finite-state modulation codes for data storage, *IEEE J. Select. Areas Commun.* **SAC-10**(1): 5–37 (Jan. 1992).
20. J. Moon and B. Brickner, Maximum transition run codes for data storage systems, *IEEE Trans. Magn.* **32**: 3992–3994 (Sept. 1996).
21. M. Oberg and P. Siegel, Performance analysis of turbo-equalized partial response channels, *IEEE Trans. Commun.* **49**(3): 436–444 (March 2001).
22. D. Palmer, J. Hong, D. Stanek, and R. Wood, Characterization of the read/write process for magnetic recording, *IEEE Trans. Magn.* **MAG-31**(2): 1071–1076 (March 1995).
23. T. Pan et al., A trellis-coded E^2PRML digital read/write channel IC, *ISSCC Digest of Tech. Papers*, Paper MP 2.2, 36–37 (Feb. 1999).
24. J. Proakis, *Digital Communications*, 2nd Edition, McGraw-Hill, New York, 1989.
25. P. H. Siegel and J. K. Wolf, Modulation and coding for information storage, *IEEE Commun. Mag.* **29**(12): 68–86 (Dec. 1991).
26. H. K. Thapar and A. M. Patel, A class of partial response systems for increasing storage density in Magnetic Recording, *IEEE Trans. Magn.* **MAG-23**(5): 3666–3668 (Sept. 1987).
27. H. Thapar, P. Ziperovich, and R. Wood, On the performance of symbol- and fractionally-spaced equalization in digital magnetic recording, *Proc. of IEEE Audio, Video, and Data Recording*, (May 1990).
28. D. A. Thompson and J. S. Best, The future of magnetic data storage technology, *IBM J. Res. Devel.* **44**(3): 311–322 (May 2000).
29. R. W. Wood and D. A. Peterson, Viterbi detection of class IV partial response on a magnetic recording channel, *IEEE Trans. Commun.* **COM-34**(5): 454–461 (May 1986).
30. R. W. Wood, Magnetic recording systems, *Proc. IEEE* **74**(11): 1557–1569 (Nov. 1986).
31. R. Wood, Turbo-PRML: A compromise EPRML detector, *IEEE Trans. Magn.* **29**: 4018–4020 (Nov. 1993).
32. *IEEE J. Select. Areas Commun.* **SAC-10**(1): (Jan. 1992).
33. *IEEE J. Select. Areas Commun.* **SAC-19**(4): (April 2001).

MATCHED FILTERS IN SIGNAL DEMODULATION

JOHN G. PROAKIS
Northeastern University
Boston, Massachusetts

1. INTRODUCTION

In digital communication systems, the modulator maps a sequence of information bits into signal waveforms that are transmitted through the communication channel. The simplest form of digital modulation is binary modulation, in which the information bit 0 is mapped into a signal waveform $s_0(t)$ and the information bit 1 is mapped by the modulator into the signal waveform $s_1(t)$. Thus, if the binary data rate into the modulator is R bits per second, each waveform may be confined to occupy a time duration $T_b = 1/R$ seconds, where T_b is called the *bit interval*. Then, the mapping performed by the modulator for binary signalling may be expressed as

$$\begin{aligned} 0 &\rightarrow s_0(t), \quad 0 \le t \le T_b \\ 1 &\rightarrow s_1(t), \quad 0 \le t \le T_b \end{aligned}$$

Higher-level modulation can be performed by mapping multiple data bits into corresponding signal waveforms. Specifically, the modulator may employ $M = 2^k$ different signal waveforms to map groups of k bits at a time for transmission through the communication channel. A group of k bits is called a *symbol*, and, for a data rate of R bits per second, the corresponding signal waveforms generally may be of duration $T_s = k/R = kT_b$ seconds. T_s is called the *symbol duration*, and the modulator for $M > 2$ is generally said to perform M-ary signal modulation.

For example, $M = 4$ signal waveforms are used to transmit pairs of data bits. The mapping performed by the modulator for $M = 4$ may be expressed as (with $T_s = 2T_b$)

$$\begin{aligned} 00 &\rightarrow s_0(t), \quad 0 \le t \le T_s \\ 01 &\rightarrow s_1(t), \quad 0 \le t \le T_s \\ 10 &\rightarrow s_2(t), \quad 0 \le t \le T_s \\ 11 &\rightarrow s_3(t), \quad 0 \le t \le T_s \end{aligned}$$

The set of signal waveforms $\{s_m(t), m = 0, 1, \ldots, M-1\}$ for $M = 2^k, k = 1, 2, \ldots$, which convey the information bits may differ either in amplitude, as in pulse amplitude modulation (PAM), or in phase, as in phase shift keying (PSK), or in both amplitude and phase, as in quadrature amplitude modulation (QAM); or, more generally, they may be multidimensional signal waveforms constructed from different frequencies, as in M-ary frequency shift keying (MFSK), or from pulses transmitted in different time slots as in M-ary time shift keying (MTSK).

In the transmission of the signal through the channel, the signal is corrupted by additive noise. This noise originates at the front end of the receiver and is well modeled statistically as a Gaussian random process. Hence, if the transmitted signal is $s_m(t), 0 \le t \le T_s$, the received signal $r(t)$ may be expressed as

$$r(t) = s_m(t) + n(t), \ 0 \le t \le T_s$$

where $m = 0, 1, \ldots, M-1$.

2. SIGNAL DEMODULATION

The basic function of the demodulation process is to recover the transmitted data by processing the received signal $r(t)$. Since the noise in the received signal in any signaling interval of duration T_s is a sample function of a random process, the demodulation of $r(t)$ should be designed to minimize the probability of a symbol error or, equivalently, to maximize the probability of a correct decision. The probability of error is the probability of selecting a signal waveform $s_j(t)$ when, in fact, waveform $s_i(t)$ was transmitted where $i \neq j$. On the basis of this criterion, the optimum demodulator, having observed $r(t)$ over the time interval $0 \le t \le T_s$, computes the M (posterior) probabilities

$$P_m \equiv P[s_m(t) \text{ was transmitted} \mid r(t), \ 0 \le t \le T_s]$$

and selects the signal waveform corresponding to the largest probability. It is shown in basic textbooks on digital communications [1–3] that the optimum demodulator obtained by applying this maximum aposterior probability (MAP) design criterion, when the transmitted signal is corrupted by additive white Gaussian noise (AWGN), consists of a parallel bank of M matched filters, where the filter impulse responses are matched to the M possible transmitted signal waveforms $s_m(t), m = 0, 1, \ldots, M-1$. The outputs of these M linear filters are sampled at the end of the signaling interval T_s, and the samples are passed to the detector which selects the largest of the M samples and performs the inverse mapping from the corresponding waveform to the $k = \log_2 M$ data bits. A block diagram of the optimum demodulator is illustrated in Fig. 1.

3. THE MATCHED FILTER

Consider a time-limited signal waveform as shown in Fig. 2a. A filter is said to be matched to the signal waveform $s(t)$ if its impulse response $h(t)$ is given as

$$h(t) = s(T - t), \ 0 \le t \le T$$

For the signal waveform shown in Fig. 2, the impulse response of the matched filter is shown in Fig. 2b.

Now, suppose that the signal $s(t)$ is the input to the filter whose impulse response is matched to $s(t)$. The output of the matched filter is given by the convolution integral

$$y(t) = \int_0^t s(\tau) h(t - \tau)\, d\tau \tag{1}$$

where $h(t) = s(T - t)$. By substituting for $h(t)$ in Eq. (1), one obtains the result

$$y(t) = \int_0^t s(\tau) s(T - t + \tau)\, d\tau \tag{2}$$

Figure 3 illustrates the filter output waveform $y(t)$.

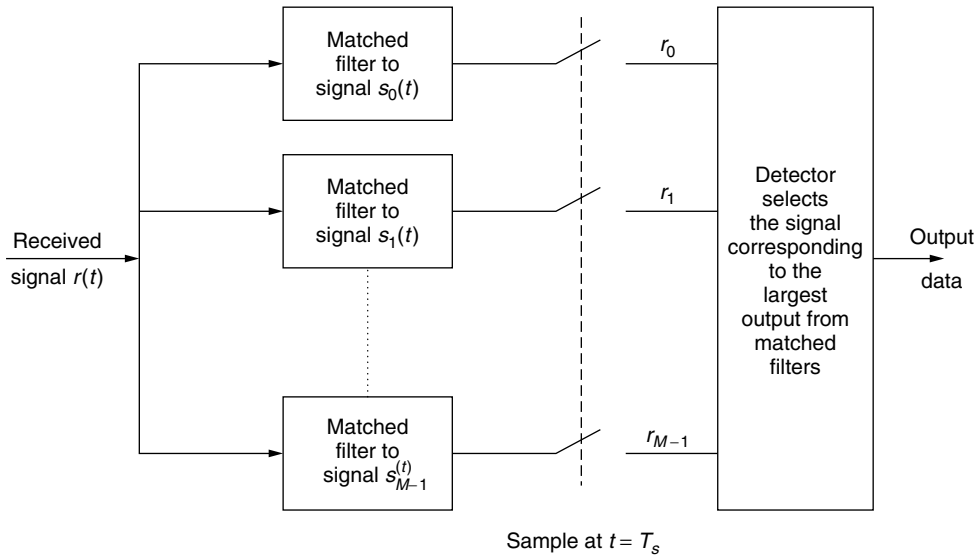


Figure 1. Signal demodulation using matched filters.

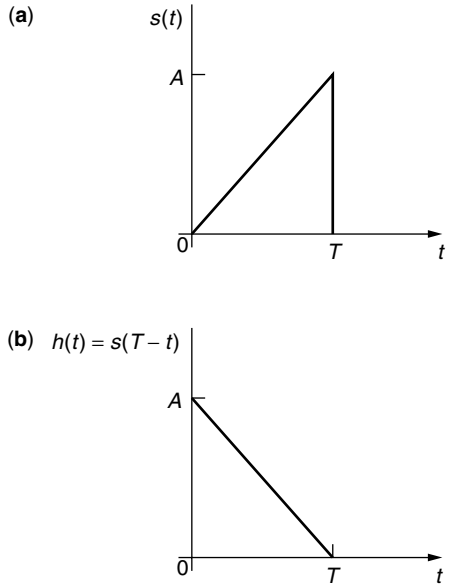


Figure 2. Signal $s(t)$ and filter matched to $s(t)$: (**a**) signal $s(t)$; (**b**) impluse response of filter matched to $s(t)$.

It is observed that $y(t)$ has a peak at $t = T$, whose value is

$$y(T) = \int_0^T s^2(\tau)\, d\tau = \mathcal{E} \tag{3}$$

which is the signal energy $\mathcal{E}$ in the signal waveform $s(t)$. Furthermore, $y(t)$ is symmetric with respect to the point $t = T$. In fact, the form of Eq. (2) is simply the time autocorrelation function of the signal $s(t)$, which is symmetric for any arbitrary signal waveform. Consequently, any signal waveform $s(t), 0 \le t \le T$, when passed through a filter matched to it, will result in an output that is the time-autocorrelation of $s(t)$, and the value of the output $y(t)$ at $t = T$ will be the energy in the signal $s(t)$, as given by equation (3).

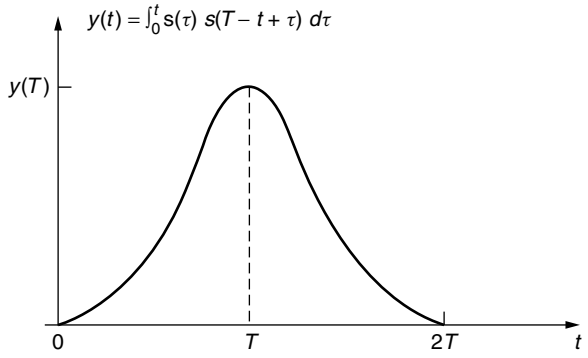


Figure 3. Output response of matched filter for the signal in Fig. 2.

4. PROPERTIES OF THE MATCHED FILTER

A matched filter has some interesting properties. We consider the most important property, which may be stated as follows: If a signal $s(t)$ is corrupted by AWGN, the filter with impulse response matched to $s(t)$ maximizes the output SNR.

To prove this property, let us assume that the received signal $r(t)$ consists of the signal $s(t)$ and AWGN $n(t)$ which has zero-mean and power-spectral density $\Phi_n(f) = N_0/2$ W/Hz. Suppose the signal $r(t)$ is passed through a filter with impulse response $h(t), 0 \le t \le T$, and its output is sampled at time $t = T$. The filter response to the signal and noise components is

$$\begin{aligned} y(t) &= \int_0^t r(\tau)h(t-\tau)\, d\tau \\ &= \int_0^t s(\tau)h(t-\tau)\, d\tau + \int_0^t n(\tau)h(t-\tau)\, d\tau \end{aligned} \tag{4}$$

At the sampling instant $t = T$, the signal and noise components are

$$\begin{aligned} y(T) &= \int_0^T s(\tau)h(T-\tau)\,d\tau + \int_0^T n(\tau)h(T-\tau)\,d\tau \\ &= y_s(T) + y_n(T) \end{aligned} \tag{5}$$

where $y_s(T)$ represents the signal component and $y_n(T)$ represents the noise component. The problem is to select the filter impulse response that maximizes the output SNR defined as

$$\left(\frac{S}{N}\right)_0 = \frac{y_s^2(T)}{E[y_n^2(T)]} \tag{6}$$

The denominator in Eq. (6) is simply the variance of the noise term at the output of the filter. Let us evaluate $E[y_n^2(T)]$. We have

$$\begin{aligned} E[y_n^2(T)] &= \int_0^T \int_0^T E[n(\tau)n(t)]h(T-\tau)h(T-t)\,dt\,d\tau \\ &= \int_0^T \int_0^T \frac{N_0}{2}\delta(t-\tau)h(T-\tau)h(T-t)\,dt\,d\tau \\ &= \frac{N_0}{2}\int_0^T h^2(T-t)\,dt \end{aligned} \tag{7}$$

Note that the variance depends on the power spectral density of the noise and the energy in the impulse response $h(t)$.

By substituting for $y_s(T)$ and $E\left[y_n^2(T)\right]$ into Eq. (6), we obtain the expression for the output SNR as

$$\left(\frac{S}{N}\right)_0 = \frac{\left[\int_0^T s(\tau)h(T-\tau)\,d\tau\right]^2}{\frac{N_0}{2}\int_0^T h^2(T-\tau)\,dt} = \frac{\left[\int_0^T h(\tau)s(T-\tau)\,d\tau\right]^2}{\frac{N_0}{2}\int_0^T h^2(T-\tau)\,dt} \tag{8}$$

Since the denominator of the SNR depends on the energy in $h(t)$, the maximum output SNR over $h(t)$ is obtained by maximizing the numerator of $(S/N)_0$ subject to the constraint that the denominator is held constant. The maximization of the numerator is most easily performed by use of the Cauchy–Schwarz inequality, which states, in general, that if $g_1(t)$ and $g_2(t)$ are finite-energy signals, then

$$\left[\int_{-\infty}^{\infty} g_1(t)g_2(t)\,dt\right]^2 \le \int_{-\infty}^{\infty} g_1^2(t)\,dt \int_{-\infty}^{\infty} g_2^2(t)\,dt$$

where equality holds when $g_1(t) = Cg_2(t)$ for any arbitrary constant C. If we set $g_1(t) = h(t)$ and $g_2(t) = s(T-t)$, it is clear that the $(S/N)_0$ is maximized when $h(t) = Cs(T-t)$; thus, $h(t)$ is matched to the signal $s(t)$. The scale factor C^2 drops out of the expression for $(S/N)_0$ since it appears in both the numerator and the denominator.

The output (maximum) SNR obtained with the matched filter is

$$\begin{aligned} \left(\frac{S}{N}\right)_0 &= \frac{2}{N_0}\int_0^T s^2(t)\,dt \\ &= \frac{2\mathcal{E}}{N_0} \end{aligned} \tag{9}$$

4.1. Frequency-Domain Interpretation of the Matched Filter

The matched filter has an interesting frequency-domain interpretation. Since $h(t) = s(T-t)$, the Fourier transform of this relationship is

$$\begin{aligned} H(f) &= \int_0^T s(T-t)e^{-j2\pi ft}\,dt \\ &= \left[\int_0^T s(\tau)e^{j2\pi f\tau}d\tau\right]e^{-j2\pi fT} \\ &= S^*(f)e^{-j2\pi fT} \end{aligned} \tag{10}$$

We observe that the matched filter has a frequency response that is the complex conjugate of the transmitted signal spectrum multiplied by the phase factor $e^{-j2\pi fT}$, which represents the sampling delay of T. In other words, $|H(f)| = |S(f)|$, so that the magnitude response of the matched filter is identical to the transmitted signal spectrum. On the other hand, the phase of $H(f)$ is the negative of the phase of $S(f)$.

Now, if the signal $s(t)$, with spectrum $S(f)$, is passed through the matched filter, the filter output has a spectrum $Y(f) = |S(f)|^2 e^{-j2\pi fT}$. Hence, the output waveform is

$$\begin{aligned} y_s(t) &= \int_{-\infty}^{\infty} Y(f)e^{j2\pi ft}\,df \\ &= \int_{-\infty}^{\infty} |S(f)|^2 e^{-j2\pi fT}e^{j2\pi ft}\,df \end{aligned} \tag{11}$$

By sampling the output of the matched filter at $t = T$, we obtain

$$y_s(T) = \int_{-\infty}^{\infty} |S(f)|^2\,df = \int_0^T s^2(t)\,dt = \mathcal{E} \tag{12}$$

where the last step follows from Parseval's relation.

The noise of the output of the matched filter has a power spectral density

$$\Phi_0(f) = |H(f)|^2\frac{N_0}{2} \tag{13}$$

Hence, the total noise power at the output of the matched filter is

$$\begin{aligned} P_n &= \int_{-\infty}^{\infty} \Phi_0(f)\,df \\ &= \int_{-\infty}^{\infty} \frac{N_0}{2}|H(f)|^2\,df = \frac{N_0}{2}\int_{-\infty}^{\infty} |S(f)|^2\,df = \frac{\mathcal{E}N_0}{2} \end{aligned} \tag{14}$$

The output SNR is simply the ratio of the signal power P_s, given by

$$P_s = y_s^2(T)$$

to the noise power P_n. Hence

$$\left(\frac{S}{N}\right)_0 = \frac{P_s}{P_n} = \frac{\mathcal{E}^2}{\mathcal{E}N_0/2} = \frac{2\mathcal{E}}{N_0} \tag{15}$$

which agrees with the result given by Eq. (9).

5. CONCLUDING REMARKS

Matched filters are widely used for signal demodulation in digital communication systems and in radar signal receivers. In the latter, the transmitted signal usually consists of a series of signal pulses. When the signal pulses are reflected from an object, such as an airplane, the received signal over the observation interval has the form $r(t) = s(t - t_0) + n(t)$, where $n(t)$ represents the additive noise and t_0 represents the round-trip time delay corresponding to the signal reflected from the object. By passing the received signal $r(t)$ through the filter matched to $s(t)$ and determining when the matched-filter output reaches a peak value that exceeds a predetermined threshold, an estimate of the time delay is obtained. From this measurement of t_0, the distance (range) of the object from the radar position is determined. If the threshold is not exceeded during an observation interval, a decision is made that no target or object is present at that corresponding range.

BIOGRAPHY

Dr. John G. Proakis received the B.S.E.E. from the University of Cincinnati in 1959, the M.S.E.E. from MIT in 1961, and the Ph.D. from Harvard University in 1967. He is an Adjunct Professor at the University of California at San Diego and a Professor Emeritus at Northeastern University. He was a faculty member at Northeastern University from 1969 through 1998 and held the following academic positions: Associate Professor of Electrical Engineering, 1969–1976; Professor of Electrical Engineering, 1976–1998; Associate Dean of the College of Engineering and Director of the Graduate School of Engineering, 1982–1984; Interim Dean of the College of Engineering, 1992–1993; Chairman of the Department of Electrical and Computer Engineering, 1984–1997. Prior to joining Northeastern University, he worked at GTE Laboratories and the MIT Lincoln Laboratory.

His professional experience and interests are in the general areas of digital communications and digital signal processing and more specifically, in adaptive filtering, adaptive communication systems and adaptive equalization techniques, communication through fading multipath channels, radar detection, signal parameter estimation, communication systems modeling and simulation, optimization techniques, and statistical analysis. He is active in research in the areas of digital communications and digital signal processing and has taught undergraduate and graduate courses in communications, circuit analysis, control systems, probability, stochastic processes, discrete systems, and digital signal processing. He is the author of the book *Digital Communications* (McGraw-Hill, New York: 1983, first edition; 1989, second edition; 1995, third edition; 2001, fourth edition), and co-author of the books *Introduction to Digital Signal Processing* (Macmillan, New York: 1988, first edition; 1992, second edition; 1996, third edition), *Digital Signal Processing Laboratory* (Prentice-Hall, Englewood Cliffs, NJ, 1991); *Advanced Digital Signal Processing* (Macmillan, New York, 1992), *Algorithms for Statistical Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 2002), *Discrete-Time Processing of Speech Signals* (Macmillan, New York, 1992, IEEE Press, New York, 2000), *Communication Systems Engineering* (Prentice-Hall, Englewood Cliffs, NJ: 1994, first edition; 2002, second edition), *Digital Signal Processing Using MATLAB V.4* (Brooks/Cole-Thomson Learning, Boston, 1997, 2000), and *Contemporary Communication Systems Using MATLAB* (Brooks/Cole-Thomson Learning, Boston, 1998, 2000). Dr. Proakis is a Fellow of the IEEE. He holds five patents and has published over 150 papers.

BIBLIOGRAPHY

1. J. G. Proakis and M. Salehi, *Communication Systems Engineering*, 2nd ed., Prentice-Hall, Upper Saddle River, NJ, 2002.
2. S. Haykin, *Communication Systems*, 4th ed., Wiley, New York, 2000.
3. H. Stark, F. B. Tuteur, and J. B. Anderson, *Modern Electrical Communication Systems*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1988.

MAXIMUM-LIKELIHOOD ESTIMATION

SIMON HAYKIN
McMaster University
Hamilton, Ontario, Canada

1. INTRODUCTION

Estimation theory is a branch of probability and statistics that deals with the problem of deriving information about properties of random variables and stochastic processes, given a set of observed data. This problem arises frequently in the study of communication and control systems. *Maximum likelihood* is a powerful method of parameter estimation, which was pioneered by Fisher [1]. In principle, the method of maximum likelihood may be applied to any estimation problem, with the proviso that we formulate the joint probability density function of the available set of observed data. The method then yields almost all the well-known estimates as special cases.

2. LIKELIHOOD FUNCTION

The method of maximum likelihood is based on a relatively simple idea:

> Different populations tend to generate different data samples, where the given data sample is more *likely* to have come from some population than from other populations.

Let $f_{\mathbf{U}}(\mathbf{u} \mid \theta)$ denote the *conditional joint probability density function* of the *random vector* $\mathbf{U}$ represented by the observed sample vector $\mathbf{u}$ with elements $u_1, u_2, \ldots, u_M$, where θ is a parameter vector with elements $\theta_1, \theta_2, \ldots, \theta_K$. The method of maximum likelihood is based on the principle that we should estimate the parameter vector θ by its most *plausible value*, given the observed sample vector $\mathbf{u}$. In other words, the maximum-likelihood

estimates of $\theta_1, \theta_2, \ldots, \theta_K$ are those values of the parameter vector for which the conditional joint probability density function $f_{\mathbf{U}}(\mathbf{u} \mid \theta)$ is a maximum.

The term *likelihood function*, denoted by $l(\theta)$, is given to the conditional joint probability density function $f_{\mathbf{U}}(\mathbf{u} \mid \theta)$, viewed as a function of the parameter vector θ. We thus write

$$l(\theta) = f_{\mathbf{U}}(\mathbf{u} \mid \theta) \tag{1}$$

Although the conditional joint probability density function and the likelihood function have exactly the same formula, it is vital that we appreciate the physical distinction between them. In the case of the conditional joint probability density function, the parameter vector θ is fixed and the observation vector $\mathbf{u}$ is variable. In the case of the likelihood function, we have the opposite situation in that the parameter vector θ is variable and the observation vector $\mathbf{u}$ is fixed.

In many cases, it turns out to be more convenient to work with the natural logarithm of the likelihood function rather than with the likelihood itself. Thus, using $L(\theta)$ to denote the *loglikelihood function*, we write

$$\begin{aligned} L(\theta) &= \ln[l(\theta)] \\ &= \ln[f_{\mathbf{U}}(\mathbf{u} \mid \theta)] \end{aligned} \tag{2}$$

The logarithmic function $L(\theta)$ is a *monotonic transformation* of $l(\theta)$. This means that whenever $l(\theta)$ decreases, its logarithm $L(\theta)$ also decreases. Where $l(\theta)$ is a formula for a conditional joint probability density function, it follows that it never becomes negative; hence there is no problem in evaluating the logarithmic function $L(\theta)$. We conclude, therefore, that the parameter vector for which the likelihood function $l(\theta)$ is a maximum is exactly the same as the parameter vector for which the loglikelihood function $L(\theta)$ is a maximum.

To obtain the ith element of the maximum-likelihood estimate of the parameter vector θ, we differentiate the loglikelihood function with respect to θ_i and set the result equal to zero. We thus get a set of first-order conditions:

$$\frac{\partial L}{\partial \theta_i} = 0, \quad i = 1, 2, \ldots, K \tag{3}$$

The first derivative of the loglikelihood function with respect to the parameter θ_i is called the *score* for that parameter. The vector of such parameters is known as the *scores vector* (i.e., the gradient vector). The scores vector is identically zero at the maximum-likelihood estimates of the parameters [i.e., at the values of parameter vector θ that result from the solutions of Eq. (3)].

To find how effective the method of maximum likelihood is, we need to compute the *bias* and *variance* for the estimate of each parameter. However, this is frequently difficult to do. Thus, rather than approach the computation directly, we may derive a *lower bound* on the variance of any *unbiased* estimate. We say an estimate is unbiased if the average value of the estimate equals the parameter we are trying to estimate. Later, we show how the variance of the maximum-likelihood estimate compares with this lower bound.

3. CRAMÉR–RAO INEQUALITY

Let $\mathbf{U}$ be a random vector with conditional joint probability density function $f_{\mathbf{U}}(\mathbf{u} \mid \boldsymbol{\theta})$, where $\mathbf{u}$ is the observed sample vector with elements $u_1, u_2, \ldots, u_M$ and θ is the parameter vector with elements $\theta_1, \theta_2, \ldots, \theta_K$. Using the definition of Eq. (2) for the loglikelihood function $L(\theta)$ in terms of the conditional joint probability density function $f_{\mathbf{U}}(\mathbf{u} \mid \boldsymbol{\theta})$, we form the $K \times K$ matrix

$$\mathbf{J} = -\begin{bmatrix} E\left[\dfrac{\partial^2 L}{\partial \theta_1^2}\right] & \left[\dfrac{\partial^2 L}{\partial \theta_1 \partial \theta_2}\right] & \cdots & E\left[\dfrac{\partial^2 L}{\partial \theta_1 \partial \theta_K}\right] \\ E\left[\dfrac{\partial^2 L}{\partial \theta_2 \partial \theta_1}\right] & E\left[\dfrac{\partial^2 L}{\partial \theta_2^2}\right] & \cdots & E\left[\dfrac{\partial^2 L}{\partial \theta_2 \partial \theta_K}\right] \\ \vdots & \vdots & & \vdots \\ E\left[\dfrac{\partial^2 L}{\partial \theta_K \partial \theta_1}\right] & E\left[\dfrac{\partial^2 L}{\partial \theta_K \partial \theta_2}\right] & \cdots & E\left[\dfrac{\partial^2 L}{\partial \theta_K^2}\right] \end{bmatrix} \tag{4}$$

The matrix $\mathbf{J}$ is called *Fisher's information matrix*.

Let $\mathbf{I}$ denote the inverse of Fisher's information matrix $\mathbf{J}$. Let I_{ii} denote the ith diagonal element (i.e., the element in the ith row and ith column) of the inverse matrix $\mathbf{I}$. Let $\hat{\theta}_i$ be *any* unbiased estimate of the parameter θ_i, based on the observed sample vector $\mathbf{u}$. We may then write [2]

$$\operatorname{var}[\hat{\theta}_i] \geq I_{ii}, \quad i = 1, 2, \ldots, K \tag{5}$$

This equation is called the *Cramér–Rao inequality*. It enables us to construct a lower limit (greater than zero) for the variance of any unbiased estimator, provided, of course, that we know the functional form of the loglikelihood function. The lower limit is called the *Cramér–Rao lower bound*.

If we can find an unbiased estimator whose variance equals the Cramér–Rao lower bound, then according to Eq. (5), there is no other unbiased estimator with a smaller variance. Such an estimator is said to be *efficient*.

4. PROPERTIES OF MAXIMUM-LIKELIHOOD ESTIMATORS

Not only is the method of maximum likelihood based on an intuitively appealing idea (that of choosing those parameters from which the actually observed sample vector is most likely to have come), but the resulting estimates also have some desirable properties. Indeed, under quite general conditions, the following *asymptotic* properties may be proved [2]:

1. Maximum-likelihood estimators are *consistent*; that is, the value of θ_i for which the score $\partial L/\partial \theta_i$ is identically zero *converges in probability* to the true value of the parameter θ_i, $i = 1, 2, \ldots, K$, as the *sample size* M approaches infinity.
2. Maximum-likelihood estimators are *asymptotically efficient*:

$$\lim_{M \to \infty} \left\{ \frac{\operatorname{var}[\theta_{i,\mathrm{ml}} - \theta_i]}{I_{ii}} \right\} = 1, \quad i = 1, 2, \ldots, K \tag{6}$$

where $\theta_{i,ml}$ is the maximum-likelihood estimate of parameter θ_i and I_{ii} is the ith diagonal element of the inverse of Fisher's information matrix.

3. Maximum-likelihood estimators are *asymptotically Gaussian*.

In practice, we find that the large-sample (i.e., asymptotic) properties of maximum-likelihood estimators hold rather well for sample size $M \geq 50$.

5. CONDITIONAL MEAN ESTIMATOR

Another classic problem in estimation theory is the *Bayes estimation of a random parameter*. There are different answers to this problem, depending on how the Bayes estimation is formulated [2]. A particular type of the Bayes estimator of interest is the *conditional mean estimator*. We now wish to do two things: (1) derive the formula for the conditional mean estimator and (2) show that such an estimator is the same as a minimum mean-square-error estimator.

Toward those ends, consider a *random parameter* x. We are given an observation y that depends on x, and the requirement is to estimate x. Let $\hat{x}(y)$ denote an *estimate* of the parameter x; the symbol $\hat{x}(y)$ emphasizes the fact that the estimate is a function of the observation y. Let $C(x, \hat{x}(y))$ denote a *cost function* that depends on both x and $\hat{x}(y)$. Let E denote the statistical expectation operator. Then according to Bayes' estimation theory, we may write the expression

$$\begin{aligned}\mathcal{R} &= E[C(x, \hat{x}(y))] \\ &= \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} C(x, \hat{x}(y)) f_{X,Y}(x, y)\, dy \end{aligned} \tag{7}$$

for the risk [2]. Here, $f_{X,Y}(x, y)$ is the joint probability density function of x and y. For a specified cost function $C(x, \hat{x}(y))$, the *Bayes estimate* is defined as the estimate $\hat{x}(y)$ that *minimizes* the risk $\mathcal{R}$.

A cost function of particular interest is the mean-square error, specified as the square of the estimation error, which is itself defined as the difference between the actual parameter value x and the estimate $\hat{x}(y)$:

$$\varepsilon = x - \hat{x}(y) \tag{8}$$

Correspondingly, the cost function is defined by

$$C(x, \hat{x}(y)) = C(x - \hat{x}(y))$$

or simply

$$C(\varepsilon) = \varepsilon^2 \tag{9}$$

Thus the cost function $C(\varepsilon)$ varies with the estimation error ε in the manner indicated in Fig. 1. It is assumed here that x and y are both real. Accordingly, we may rewrite Eq. (7) for mean-square error as

$$\mathcal{R}_{\text{ms}} = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} [x - \hat{x}(y)]^2 f_{X,Y}(x, y)\, dy \tag{10}$$

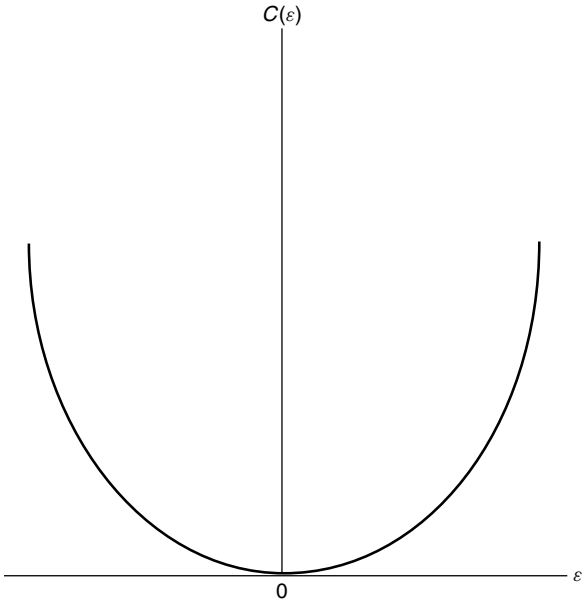


Figure 1. Mean-square error as a quadratic cost function.

where the subscripts ms in the risk $\mathcal{R}_{\text{ms}}$ indicate the use of mean-square error estimation. From probability theory, we have

$$f_{X,Y}(x, y) = f_X(x \mid y) f_Y(y) \tag{11}$$

where $f_X(x \mid y)$ is the conditional probability density function of x given y, and $f_Y(y)$ is the (marginal) probability density function of y. Hence, using Eq. (11) in Eq. (10), we may write

$$\mathcal{R}_{\text{ms}} = \int_{-\infty}^{\infty} dy f_Y(y) \int_{-\infty}^{\infty} [x - \hat{x}(y)]^2 f_X(x, y)\, dx \tag{12}$$

We now recognize that the inner integrand and the probability density function $f_Y(y)$ in Eq. (12) are both nonnegative. We may therefore simplify matters by minimizing the inner integral. Let the estimate so obtained be denoted by $\hat{x}_{\text{ms}}(y)$. We find $\hat{x}_{\text{ms}}(y)$ by differentiating the inner integral with respect to $\hat{x}(y)$ and then setting the result equal to zero. To simplify the minimization procedure, let I_{inner} denote the inner integral in Eq. (12). Then differentiating I_{inner} with respect to the estimate $\hat{x}(y)$ yields

$$\frac{dI_{\text{inner}}}{d\hat{x}} = -2 \int_{-\infty}^{\infty} x f_X(x \mid y)\, dx + 2\hat{x}(y) \int_{-\infty}^{\infty} f_X(x \mid y)\, dx \tag{13}$$

The second integral on the right-hand side of Eq. (13) represents the total area under a probability density function, which, by definition, equals unity. Hence, setting the derivative $dI_{\text{inner}}/d\hat{x}$ equal to zero and solving for the minimum mean-square error estimate, denoted by, we obtain

$$\hat{x}_{\text{ms}}(y) = \int_{-\infty}^{\infty} x f_X(x \mid y)\, dx \tag{14}$$

The optimum solution defined by Eq. (14) is unique by virtue of the assumed form of the cost function $C(\varepsilon)$. For another interpretation of the estimator $\hat{x}_{\text{ms}}(y)$, we

recognize that the integral on the right-hand side of the equation is just the *conditional mean* of the parameter x, given the observation y. On the basis of this result, we therefore conclude that *the minimum mean-square-error estimate and the conditional mean estimator are indeed one and the same*. In other words, we have

$$\hat{x}_{\text{ms}}(y) = E[x \mid y] \tag{15}$$

Substituting Eq. (15) for the estimate $\hat{x}(y)$ into Eq. (12), we find that the inner integral is just the conditional variance of the parameter x, given y. Accordingly, the minimum value of the risk $\mathcal{R}_{\text{ms}}$ is just the average of this conditional variance over all observations y.

6. EXPECTATION–MAXIMIZATION (EM) ALGORITHM

The *expectation–maximization algorithm*, popularly known as the *EM algorithm*, is an iterative algorithm for computing maximum-likelihood estimates when dealing with data that have a latent structure and/or are incomplete. Moreover, computation of the maximum-likelihood estimate is often greatly facilitated by formulating it as an incomplete data problem, which is invoked because the EM algorithm is able to exploit the reduced complexity of the maximum-likelihood estimate, given the complete data. Applications of the EM algorithm include hidden Markov models for speech recognition and hierarchical mixture of experts model for the design of neural networks.

The EM algorithm derives its name from the fact that, at each iteration of the algorithm, there are two basic steps [3,4]:

- *Expectation step* or *E-step*, which uses the given data set of an incomplete data problem and the current value of the parameter vector to manufacture data so as to postulate an augmented or so-called complete data set.
- *Maximization step* or *M-step*, which consists of deriving a new estimate of the parameter vector by maximizing the loglikelihood function of the complete data manufactured in the E-step.

The E-step, operating in the forward direction, and the M-step, operating in the backward direction, form a closed loop. Thus, starting from a suitable value for the parameter vector, the E-step and M-step are repeated on an alternating basis until convergence occurs.

Let the vector $\mathbf{z}$ denote the missing or hidden data. Let $\mathbf{r}$ denote the complete data vector, made up of some observable data d and the missing data vector $\mathbf{z}$. There are therefore two data spaces $\mathcal{R}$ and $\mathcal{D}$ to be considered, and the mapping from $\mathcal{R}$ to $\mathcal{D}$ is many-to-one. However, instead of observing the complete data vector $\mathbf{r}$, we are actually able to observe only the complete data $d = d(\mathbf{r})$ in $\mathcal{D}$. Let $f_c(\mathbf{r} \mid \theta)$ denote the conditional probability density function (pdf) or $\mathbf{r}$, given a parameter vector θ. It follows therefore that the conditional PDF of random variable D, given θ, is defined by

$$f_D(d \mid \theta) = \int_{\mathcal{R}(d)} f_c(\mathbf{r} \mid \theta)\, d\mathbf{r} \tag{16}$$

where $\mathcal{R}(d)$ is the subspace of $\mathcal{R}$ that is determined by $d = d(\mathbf{r})$. The EM algorithm is directed at finding a value of θ that maximizes the *incomplete data loglikelihood function*

$$L(\theta) = \log f_D(d \mid \theta) \tag{17}$$

This problem, however, is solved indirectly by working iteratively with the *complete data loglikelihood function*

$$L_c(\theta) = \log f_c(\mathbf{r} \mid \theta) \tag{18}$$

which is a random variable, because the missing data vector $\mathbf{z}$ is unknown.

To be more specific, let $\hat{\theta}(n)$ denote the value of the parameter vector θ on iteration n of the EM algorithm. In the E-step of this iteration, we calculate the expectation

$$Q(\theta, \hat{\theta}(n)) = E[L_c(\theta)] \tag{19}$$

where the expectation is performed with respect to $\hat{\theta}(n)$. In the M-step of this same iteration, we maximize $Q(\theta, \hat{\theta}(n))$ with respect to θ over the parameter (weight) space $\mathcal{W}$, and so find the updated parameter estimate $\hat{\theta}(n+1)$, as shown by

$$\hat{\theta}(n+1) = \arg\max_{\theta} Q(\theta, \hat{\theta}(n)) \tag{20}$$

The algorithm is started with some initial value $\hat{\theta}(0)$ of the parameter vector θ. The E-step and M-step are then alternately repeated in accordance with Eqs. (19) and (20), respectively, until the difference between $L(\hat{\theta}(n+1))$ and $L(\hat{\theta}(n))$ drops to some arbitrary small value; at that point the computation is terminated. Note that after an iteration of the EM algorithm, the incomplete data loglikelihood function is *not* decreased, as shown by

$$L(\hat{\theta}(n+1) \geq L\hat{\theta}(n)) \quad \text{for } n = 0, 1, 2, \ldots \tag{21}$$

Equality usually means that we are at a stationary point of the loglikelihood function.

Under fairly general conditions, the loglikelihood function computed by the EM algorithm converges to stationary values. However, a cautionary note is in order. The EM algorithm will not always lead to a local or global maximum of the loglikelihood function. This point is demonstrated in Chapter 3 of the book by McLachlan and Krishnam [4]; in one of two examples presented therein, the EM algorithm converges to a saddle point, and in the other example the algorithm converges to a local minimum of the loglikelihood function.

7. DISCUSSION

In this article, we presented a description of maximum-likelihood (ML) estimation, which, in mathematical terms, corresponds to the limiting case of maximum a posteriori probability (MAP) estimation when the a priori knowledge pertaining to the problem at hand approaches zero. ML estimates have some nice asymptotic properties as the size of the data set approaches infinity. Indeed, it is these properties that motivate the use of ML estimates even when there is no efficient estimate.

We also briefly described a forward–backward computation procedure known as the EM algorithm, which is

remarkable in part because of the simplified and generality of the underlying theory, and in part because of the wide range of applications that fall under its umbrella. The EM algorithm applies to incomplete data problems. Problems of this kind encompass situations where naturally there are hidden variables, and other situations where the incompleteness of data is not at all evident or natural to the problem of interest.

BIOGRAPHY

Simon Haykin received the degrees of B.Sc. (First Class Honours), Ph.D., and D.Sc., all in electrical engineering from the University of Birmingham, England. On the completion of his Ph.D. studies, he spent several years from 1956 to 1965 in industry and academe in England. In January 1966, he joined McMaster University, Hamilton, Ontario, Canada, as Full Professor of Electrical Engineering; he has stayed there since. In 1996, the Senate of McMaster University established the new title of University Professor; in April of that year, he was appointed the first University Professor from the Faculty of Engineering.

Professor Haykin is a Fellow of the IEEE and a Fellow of the Royal Society of Canada. In 1999 he was awarded the honorary degree of Doctor of Technical Sciences by ETH, Zurich, Switzerland.

Professor Haykin's research interests have focused on adaptive signal processing, for which he is recognized worldwide.

BIBLIOGRAPHY

1. R. A. Fisher, Theory of statistical estimation, *Proc. Cambridge Phil. Soc.* **22**: 700–725 (1925).
2. H. L. Van Trees, *Detection, Estimation, and Modulation Theory*, Part I, Wiley, 1968.
3. A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Stat. Soc., B* **39**: 1–38 (1977).
4. G. J. McLachlan and T. Krishnam, *The EM Algorithm and Extensions*, Wiley, 1997.

MEDIUM ACCESS CONTROL (MAC) PROTOCOLS

ANDRÁS FARAGÓ
VIOLET R. SYROTIUK
University of Texas at Dallas
Richardson, Texas

1. INTRODUCTION

A number of communication networks use a *broadcast* (or *multiaccess*) *transmission medium*, where the signal transmitted by a node (station) is received by every other node that is in the listening area of the transmitting node. The most frequently occurring examples of such networks are wired or wireless local-area networks (LANs), and radio networks that include, as examples, satellite networks, wireless cellular networks, and mobile ad hoc radio networks (also called *packet radio networks*).

Because of the nature of the broadcast medium and the technological constraints of network nodes, the transmissions have to be controlled to ensure successful communication. Examples of typical technological constraints are (1) a node cannot both transmit and receive at the same time; and (2) a node can receive only one transmission at a time — in case of more than one simultaneous transmission, a *collision* occurs and nothing is received successfully. The constraints vary by technology and also by network type. The general task of the *medium access control* (MAC) *protocol* is to organize the transmissions such that under the given constraints successful communication takes place over the broadcast transmission medium.

The MAC protocol is fundamental to the ability to communicate in, and the performance of, networks based on broadcast channels. It is thus a vast and well-studied topic because of the wide range of broadcast media and the importance of the task. A large number of MAC protocols have been proposed in the literature, and quite a few of them are standardized and widely deployed in different commercial networks. In the next section we classify MAC protocols according to various criteria. Then, in subsequent sections, we provide a more detailed description of the most important protocols, categorized according to the way they implement the multiaccess communication.

[*Remark*: In point-to-point (rather than broadcast) communication-based networks, similar protocols are used for *multiplexing* several signals onto the same link. In this article, however, we discuss only MAC protocols and do not address multiplexing protocols.]

2. CLASSIFICATION OF MAC PROTOCOLS

MAC protocols can be classified by a number of different criteria, which we overview below. Details of the most important protocols are then given in the following sections.

2.1. Classification by Physical Domain of Sharing

2.1.1. Time Domain. Here the transmissions are organized in time, using the same frequency band. The time may be either *slotted* or *unslotted*. In slotted time, time is discrete and packet transmission can begin only at slot boundaries; it assumes that the nodes are synchronized, which is itself a challenging problem in a distributed setting. In unslotted or continuous time, there is no restriction on when packet transmission can begin. The key problem is how to organize the transmissions, such that collisions are either avoided or resolved with repeated transmissions.

2.1.2. Frequency Domain. If different stations can use different frequencies, then it is possible to separate the transmissions in the frequency domain by filtering or by tuning to a given frequency band. A traditional example is radio broadcasting.

2.1.3. Hybrid Domains. In several networks the physical domain of sharing is a combination of time and frequency. The most important group of hybrid techniques is *spread-spectrum* communication, commonly referred to as *code-division multiple access* (CDMA). Here we can symbolically say that the "code domain" is shared. Another example is the MAC protocol in the European digital cellular radio *Global System for Mobile Communication* (GSM) standard that combines a complex slotted timeframing hierarchy within 124 frequency channels [1].

2.2. Classification by Method of Sharing

2.2.1. Allocation-Based Protocols. In this category the transmission rights are allocated in advance to the nodes to avoid collisions. A typical example is *time-division multiple access* (TDMA) with a fixed assignment of time slots to users that they can use for transmission. If there is a change in the network topology or traffic pattern, reallocation of the slots may take place, but the standard operation is based on pre- or reallocated transmission rights.

2.2.2. Contention-Based Protocols. In these protocols the operation is fully distributed and there are no preallocated transmission rights. The stations *contend* for transmission, attempting it whenever needed. As a consequence, *collisions* may occur that make retransmission necessary. The key part of the protocol is how *collision resolution* is done to ensure that with appropriately organized retransmissions eventually all users have acceptable throughput and delay. A popular example is the IEEE 802.11 Wireless LAN protocol.

2.2.3. Hybrid Schemes. Many protocols exhibit both allocation- and contention-based features. For example, in *reservation*-based protocols there is a contention phase for reserving the channel. Once the reservations are made, the data are then transmitted as if the access were allocation-based.

2.3. Classification by Mode of Operation

2.3.1. Centralized Protocols. These protocols require a central entity that controls the channel access of all nodes in a given area. For example, in a cellular network the base station in a cell plays this role; in a satellite network the satellite provides control in its coverage area.

2.3.2. Distributed Protocols. If no central entity is available or desired, then the operation of the protocol must be distributed. This is necessary, for example, in mobile ad hoc networks and in most LANs. Contention based protocols typically operate in a distributed fashion.

2.4. Classification by Network Characteristics

The MAC protocols that are practically applied are tailored to the type of network in which they are used. Some typical factors that depend on the network characteristics are listed below.

2.4.1. Connectivity. In most wired LANs the network is logically fully connected; that is, every transmission is received by all stations. On the other hand, mobile wireless ad hoc networks have lower connectivity, which introduces new problems such as the *hidden-* and *exposed-terminal* problems (see Fig. 1). The hidden-terminal problem occurs when the destination B of a transmitting node A suffers a collision because of an interfering transmission from another node C that is not in the range of A. In this case, C is hidden from A. In the exposed-terminal problem, if node B is transmitting to A, C can transmit concurrently with B as long as its destination is not in the overlapping transmission range of B and C (e.g., node D). However, without additional information, node C cannot make this determination.

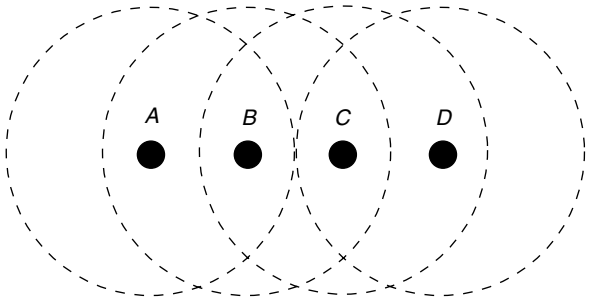


Figure 1. Each large circle indicates the transmission range of the node at its center.

2.4.2. Propagation Time. The performance of the MAC protocol is seriously influenced by the relationship of the propagation delay D and the packet transmission time T. This is often expressed by the ratio D/T. In most LANs, mobile ad hoc networks and cellular networks the ratio is small, around 10^{-2}. This implies that a collision can be quickly detected. On the other hand, in satellite networks the ratio can be as large as 100, so many packets may be sent before a collision is detected.

2.4.3. Physical Link Characteristics. Wireless links behave substantially differently from wired links. The signal-to-noise ratio is lower, while the bit error rate and signal attenuation are typically higher in wireless links. This also has an influence on optimizing the MAC protocol for a given network. An example is the *near–far* problem when the stronger signal of a node nearby the destination suppresses the weaker signal of a remote station. Instead of a collision being detected at the receiver, the closer node "captures" the receiver. As a result, the remote node may not be helped out by collision resolution.

2.5. Summary

As seen above, there are a number of ways to categorize the large number of existing MAC protocols. In the subsequent sections we review some of the most important MAC protocols, according to their method of sharing the medium (allocation- or contention-based, or a hybrid of these). We choose this categorization for our presentation because the most characteristic aspect of a MAC protocol is how it actually implements the sharing of the multiaccess medium.

The traditional performance measures of a MAC protocol are its throughput and delay characteristics at

various traffic load conditions. In this brief overview of MAC protocols, there is no opportunity to introduce the models and develop the mathematical foundations for the performance analysis of each protocol. The interested reader should consult the book by Bertsekas and Gallager [2] for a good introduction to the analysis of MAC protocols.

3. ALLOCATION-BASED MAC PROTOCOLS

In allocation-based MAC protocols parts of the communication resources (such as time or frequency) are assigned in advance to the stations in a way that excludes collision, so there is no contention for transmission. Below we discuss some of the most important allocation based MAC protocols.

3.1. Time-Division Multiple Access (TDMA)

In TDMA protocols time is slotted and each slot can incorporate the transmission of one packet. The key issue is how to allocate the slots to stations, such that no collision occurs. Typically, two constraints have to be satisfied: (1) a node cannot both transmit and receive in the same slot (primary conflict); and (2) no successful reception is possible if more than one transmission reaches the node in a slot (secondary conflict).

If the network is logically fully connected, that is, if each transmission is received by all stations, then the allocation is conceptually very simple — the time slots are grouped into frames and each node has its own unique slot in each frame. The frames are periodically repeated, so if there are N nodes, then each one has the opportunity to transmit once in each frame of N slots.

The above mentioned simple frame structure assumes that we want to give all nodes an equal chance to transmit. If, however, different nodes generate different amounts of traffic, then they may be assigned a different number of slots in a longer frame. The assignment is called *fair* if each node is assigned a number of slots proportional to its traffic rate, that is, the average number of packets per unit time that the node wants to transmit. On the other hand, if packets are generated randomly, some slots may remain unused at certain nodes, while at others the buffer may overflow. Thus, one may also want to assign the slots so that the *throughput*, that is, the average number of successful packet transmissions per slot, is maximized. It is interesting that one can prove under general modeling assumptions the maximum throughput is achieved precisely when the assignment is fair [3].

The slot assignment problem is more complicated in mobile ad hoc networks, where the network topology is not fully connected and furthermore, can change over time. In these networks it is possible that two nodes that are more than two hops away in the network topology can be assigned the same time slot without the danger of any conflict. (Here, a "hop" refers to nodes within direct transmission range of a node — since all nodes are not one hop away from each other in a mobile ad hoc network it is often called a *mobile multihop network*.) This makes it possible to achieve *spatial reuse* of the available spectrum. An example of such a conflict-free slot assignment for the multihop network is shown in Fig. 2. Notice that the shorter frame lengths must inter-operate with longer frame lengths.

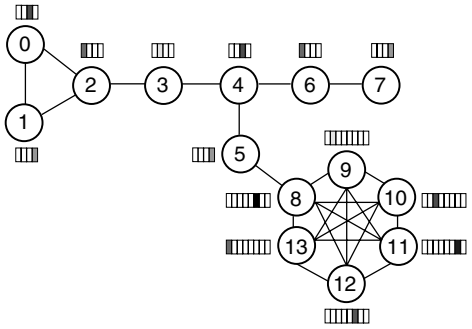


Figure 2. A conflict-free slot assignment for a multihop network.

A number of algorithms were proposed to find good slot assignments for spatial reuse TDMA [4,5]. The task can be mathematically modeled by a *graph coloring* problem where nodes of the same color can transmit concurrently. This is an algorithmically difficult (NP-complete) problem even for the restricted graphs that can occur as mobile ad hoc network topologies [6]. Nevertheless, acceptable solutions can be found with simple heuristics, such as given by a greedy algorithm.

The advantage of TDMA protocols is that they guarantee a certain throughput via the allocated transmission rights. On the other hand, for low traffic loads, TDMA introduces unnecessary delays, since a station has to wait for its turn even if others are not transmitting.

3.2. Frequency-Division Multiple Access (FDMA)

FDMA assigns a different frequency to each station. Since this makes it possible to separate the transmissions in the frequency domain, they do not have to be separated in time. In its pure form FDMA is best suited for analog systems. It has been used for a long time in radio and TV broadcasting. Note that broadcasting involves spatial reuse, since beyond the coverage area of a radio station its frequency can be reused. The way FDMA is used in broadcasting can only serve relatively few transmitters, as the radio spectrum is a scarce resource. FDMA is a component in a number of MAC protocols in cellular telephony, such as in the GSM system, which combines TDMA with FDMA [1].

3.3. Code-Division Multiple Access (CDMA)

One basic form of CDMA is *frequency hopping* (FH/CDMA). In this system there are a number of channels on different frequencies and the transmitter quickly hops between the different channels in a pseudo-random manner, as illustrated in Fig. 3. In this way the signal energy is spread over a larger frequency band, hence this technique is also called *spread-spectrum* communication. The spreading takes place according to a *spreading code* that specifies the hopping schedule. If the receiver "knows" the spreading code, then it can successfully receive the signal by following the same hopping schedule among the channels. Without the correct code, however, only noise is received, since the

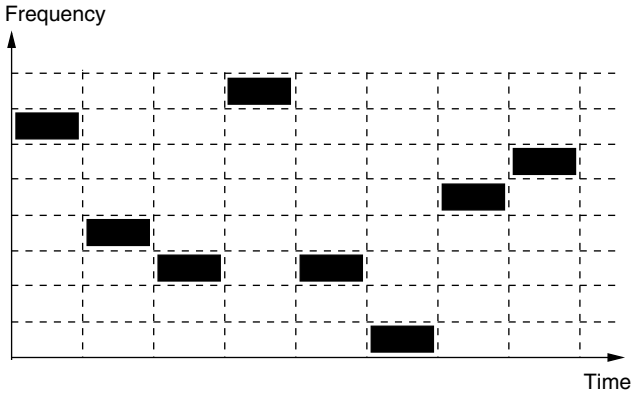


Figure 3. Frequency-hopping code-division multiple access (FH/CDMA).

signal is hidden in a broad frequency band in which the overall noise may be stronger than the useful signal that always falls into a narrow band.

Another basic form of CDMA is *direct-sequence* CDMA (DS/CDMA). Here each bit is replaced by a pseudo-random sequence of bits, which now represents the code (sometimes called a *chip sequence*). The receiver correlates the received sequence with the known code. If it is the same code, then the correlation is high; otherwise it is small (in case of orthogonal codes, it is zero). For example, assume that nodes A, B, and C are assigned the following codes, respectively: $(+1, +1, -1, -1)$; $(+1, +1, +1, +1)$; $(+1, -1, +1, -1)$. To transmit a binary one (1), a node transmits its code c; otherwise it transmits the negation of its code $\overline{c}$. When two or more nodes transmit simultaneously, their codes add linearly. To recover the transmission of a specific node, a receiver computes the normalized inner product of that node's code with the incoming signal. Since all of code pairs are orthogonal, all signals except that from the specific node are eliminated. For example, suppose that node D wants to recover the transmission from node C, and that node A transmits a zero and nodes B and C transmit ones at the same time. Node D then computes

$$C \cdot (\overline{A} + B + C) = C \cdot \overline{A} + C \cdot B + C \cdot C = 0 + 0 + 1$$

and thus node D recovers that node C transmitted a 1.

In both CDMA systems the appropriate choice of codes makes it possible for a receiver to lock onto a transmission even if other transmissions are present. This is due to the fact that, in case of orthogonal codes, the correlation with the interfering transmissions is zero, so they are effectively filtered out from the aggregated received signal. The number of orthogonal codes, however, is limited by the available spectrum. The number of codes can be significantly increased if we do not insist on full orthogonality. In this case the correlation with the interfering transmission will not be zero, but can be kept small as long as there are not too many interfering transmissions. In any case, since the codes are allocated in advance, CDMA is an allocation-based protocol, at least in its basic forms.

CDMA has several advantages. It is resistant to jamming and interception, which makes it desirable for tactical applications.[1] With appropriately designed codes, it can work without global synchronization. A characteristic feature is that CDMA has no hard limit on capacity, since the increasing number of interfering packets results in degrading signal-to-noise ratio, but not in sudden breakdown. In this way the network shows *graceful degradation* in case of increasing traffic load. In CDMA cellular networks frequency planning is easy, since each cell uses the same frequency band. Handoff is also easier and graceful degradation is an attractive feature. On the other hand, CDMA has some disadvantages, too. Its implementation technology is complex in addition to requiring power control and a large contiguous frequency band.

IS-95 CDMA is a digital cellular radio system that is used in over 35 countries worldwide. Some CDMA systems operate in the personal communications systems (PCS) frequency band.

3.4. Centralized and Distributed Polling, Token Ring

A simple allocation-based MAC solution is *polling*, when a master station polls each node in a round-robin manner to see whether it has a packet to transmit. The node that is being polled can send the packet to the master or directly to another node. This is similar to the sharing philosophy of TDMA in a fully connected network. The essential difference is, however, that if a station has nothing to send, then there is no need to wait; the master can immediately poll the next station. The simple round-robin polling scheme can be improved at the price of added complexity. There are more sophisticated polling strategies that perform a logarithmic search for a station with a packet to send (see Ref. 2, Section 4.5.6).

Polling is a centralized protocol that is also allocation-based; since the master station allocates the transmission rights, no contention is involved. To create a distributed version of polling, one can observe that it is not the master station that is essential in the protocol. What is really needed is the rotating transmission right that goes from station to station in a cyclic manner, and if a node has nothing to send, then the transmission right is quickly passed to the next station. This is the core idea of the *token-ring* protocol. In a token-ring network the nodes are arranged in a ring. Conceptually, the operation can be described such that there is a rotating token in the network that is passed from node to node. Only the station that currently has the token is allowed to transmit a packet, after which point it passes the token to the next node. If the station has no packet to transmit, then it passes the token immediately. This operation achieves, conceptually, the same effect as round-robin polling, but in a distributed way.

[1] Let us cite an interesting historical remark from Ref. 7, p. 168: "Spread spectrum (using frequency hopping) was invented, believe it or not, by Hollywood screen siren Hedy Lamarr in 1940 at the age of 26. She and a partner who later joined her effort were granted a patent in 1942 [U.S. Patent 2,292,387; Aug. 11, 1942]. Lamarr considered this her contribution to the war effort and never profited from her invention."

Of course, an implementation of the token ring has to pay attention to a number of practical issues, such as what information is put in the token (a type of control packet), how to recover from a token loss or failure, and how to recover from a node failure. The details of the protocol are standardized in the IEEE 802.4 (token bus) and IEEE 802.5 (token ring) standards [8]. In the token bus protocol, the nodes are logically arranged in a ring using a distributed algorithm to establish and maintain the ring. The token ring, on the other hand, is a physical ring, although it is common to wire the nodes to a wire center to permit the ring to remain functional in the presence of node failures or maintenance.

A higher-performance fiberoptic token ring is *Fiber Distributed Data Interface* (FDDI). The primary difference between FDDI and IEEE 802.5 is that the token is put on the ring immediately after the transmission of a packet rather than after the source drains its own packet from the ring. Thus, in FDDI, it is possible to have multiple packets on the ring at the same time, resulting in higher throughput in the presence of higher transmission rates and larger distances. FDDI is used primarily in backbone networks.

IEEE 802.4, IEEE 802.5, and FDDI are all capable of handling several priority classes of traffic with guaranteed throughput and delay [2], something that is not possible in contention based protocols.

4. CONTENTION-BASED MAC PROTOCOLS

While allocation-based MAC protocols have the advantage that they can guarantee a certain throughput and, therefore, can prevent complete breakdown of the network due to congestion, they are not very efficient for light-traffic-load situations. If the traffic load is light, a node, rather than waiting for its turn, can attempt transmission immediately when it has a packet to transmit. This is how contention-based protocols operate. Even though this savings in delay can cause collision, if the traffic load is light, the probability of collision is low. In case a collision still occurs, the protocol resolves it by resending the packet later, possibly trying multiple times. The heart of a contention based MAC protocol is in how it implements collision resolution.

Another important advantage of contention-based protocols is that they can automatically adapt to changing network topology in ad hoc networks, as opposed to TDMA that may need networkwide frame and slot reassignment whenever there is a change in the topology. On the other hand, contention protocols do not guarantee a deterministically bounded delay. Below we review some of the fundamental contention based MAC protocols.

4.1. ALOHA and Its Variants

ALOHA is the historically first contention-based MAC protocol [9], an influential milestone in MAC protocol history. It has slotted and unslotted versions, depending on whether packet transmission can be started only at time-slot boundaries or at any time, respectively. Let us explain the operating principle through the slotted version (the unslotted version is conceptually similar, only that a packet is vulnerable to collision for longer time periods).

When a node has a packet to transmit, it simply transmits it in the first available time slot. If the transmission is successful (which the node may know, e.g., from an acknowledgment sent on a separate channel or piggybacked onto a response), then there is nothing else to do with respect to that packet. If, however, the transmission is not successful, such as when a collision or transmission error occurs, then the node retransmits the packet with a *random delay*, that is, after waiting a random number of slots. In other words, the node backs off and then tries the transmission again after a random waiting time. Nothing excludes that the packet collides on successive transmission attempts, but this has lower probability, given that the network is not overloaded. Ultimately, after a number of retransmissions the packet has a very good chance to get through.

A important issue is to decide how to draw the random delay, that is, what should be the *backoff scheme*. A simple variant is when transmission occurs in each slot with a given probability p. This is called p-persistent ALOHA, and it corresponds to a geometrically distributed random delay. Another popular scheme is *binary exponential backoff*, where the next transmission slot is drawn uniformly at random from an interval and after each unsuccessful trial the length of the interval is doubled. In real networks, such as Ethernet, if the interval reaches a maximum length (1024), it remains fixed at that length. If the transmission is still unsuccessful at the maximum length, after some number (16) of collisions, failure is reported to and handled by the higher-layer protocol.

The analysis of various backoff schemes has been the subject of intense research, and it is not easy to determine which is the best one. For example, despite its wide usage, it is known that binary exponential backoff results in an unstable protocol (infinitely growing queues) under certain modeling assumptions, such as infinite user population [10]. The existence of stable protocols in this setting also depends on the type of feedback available from the channel and on how the user population is modeled. For acknowledgment-based protocols, it is known [11] that a large class of backoff, schemes, including polynomial backoff, is unstable in the infinite user population model (in polynomial backoff the backoff interval grows according to a polynomial function rather than an exponential function). In contrast, for a finite user population, any superlinear polynomial backoff protocol has been proved stable, while binary exponential backoff still remains unstable above a certain arrival rate [12].

4.2. CSMA and Its Variants, CSMA/CD and CSMA/CA

A natural improvement of ALOHA is possible if the stations can sense before transmission whether the channel is idle. After *carrier sensing*, a station starts transmitting only if the channel is idle, since otherwise the packet would surely collide with the ongoing transmission. This is the basis for the *Carrier Sense Multiple Access* (CSMA) protocol, which otherwise operates similarly to ALOHA. If the network had zero propagation delay, then collision could occur only if two nodes start transmission of a packet at precisely the same time. This has zero probability in continuous time. With finite propagation

delay, however, a node may sense that the channel is idle even if another node has already started transmitting, but due to the propagation delay, the signal has not reached yet the first node.

Further improvement to CSMA is possible via *collision detection* (CD); if a transmitting node detects collision, then it stops transmitting, since the rest of the packet transmission time is just wasted. CSMA/CD with binary exponential backoff is the basis for the MAC protocol in the IEEE 802.3 LAN standard [8]. This standard, popularly known as *Ethernet*, is by far the most widely used protocol in LANs today. Because of its widespread success, Fast Ethernet (IEEE 802.3u) did not change the protocol; it only made it run faster, with the faster technology. Another variation of CSMA uses *collision avoidance* rather than collision detection. CSMA/CA is commonly used in wireless networks since collision detection is not commonly available in wireless nodes. (Collision detection requires that a node can both transmit and receive simultaneously.) We discuss CSMA/CA in Section 5.1.

4.3. Splitting Algorithms

A more sophisticated collision resolution technique with higher theoretical performance is implemented in another class of protocols, called *splitting algorithms* (see Ref. 2, Section 4.3). The basic principle is described as follows. If in slot k a collision occurs, then the stations that are not involved in the collision go into a waiting mode. The involved nodes split into two roughly equally sized groups, for example, by drawing a random bit ("flipping a coin"). The first group attempts retransmission in slot $k+1$, while the second group in slot $k+2$. If there is no new collision (because only half of the nodes are involved), then the collision has been resolved. If there is a new collision in slot $k+1$, then all involved nodes hear it and then the first group starts resolving it using the same protocol recursively for the halved group. During this time the second group waits, since it can sense the collisions. After the first group's collisions have been resolved, the second group runs the same algorithm for itself, again recursively for a group that is half that of the original. In the worst case, the splitting continues until the group contains only one member, in which case no collision is guaranteed in the next slot. This special example of splitting protocols is called a *tree algorithm*.

5. HYBRID SCHEMES

Quite a few MAC protocols try to combine the advantages of allocation and contention. In Sections 5.1 and 5.2 we review a few interesting solutions.

5.1. Reservation and Collision Avoidance

If the network has a master station that can coordinate the operation of the other stations, such as in satellite and cellular networks, then a good way of getting rid of the wasted bandwidth caused by collisions is to make *reservations* with the master station. A typical solution is to dedicate a certain time period for making the reservation requests, and then the stations to which the master grants reservation can send their data without collision in an allocated period of time. There are different ways to make the reservations. For example, in the *Packet Demand Assignment Multiple Access* (PDAMA) scheme the stations contend for *reservation minislots* using slotted ALOHA. Then the master computes a transmission schedule and announces it to all stations [13]. This is clearly a combination of contention and allocation, where the contention phase is restricted to the reservation minislots that take only a small percentage of time. Since reservation messages are short, if the number of stations is not too large, contention can be fully eliminated, as in the *Fixed Priority Oriented Demand Assignment* (FPODA) protocol, where each station is assigned its own reservation minislot [13]. The reservation concept goes back to the beginnings of MAC protocol development, as *Reservation ALOHA* (*R-ALOHA*) was already proposed in the early 1970's [14]. Since then, many variants have been used in satellite and cellular systems [15]. Notice that even spatial reuse TDMA protocols, described in Section 3.1 use a contention period to recompute TDMA schedules in the presence of mobility.

If the network has no central master station or if it is not fully connected, reservation is not feasible. Then the *collision avoidance* (CA) technique can be applied to reduce the chance for collision. In the CSMA/CA protocol each station is assigned a time called an *interframe spacing* (IFS), which is related to the propagation delay. Additionally, lower-priority nodes are assigned a longer IFS. Before a station transmits, it waits for an IFS time. If a higher-priority station wanted to transmit simultaneously, then the lower priority station, because of its longer IFS, can already sense that the channel is busy and refrains from transmission, so collision in this case can be avoided. After waiting an IFS time the station sets a contention timer to a random value and starts counting down. When the timer expires, transmission is attempted. If during the countdown the node senses that another node transmits a packet, then the timer is frozen until the packet is completed and then the countdown continues.

In mobile ad hoc networks additional difficulties arise, due to the irregular topology, such as the *hidden-terminal problem* mentioned in Section 2. A possible solution is provided by the *Busy-Tone Multiple Access* (BTMA) protocol [16], where a node while receiving a packet transmits another signal ("busy tone") on a separate channel, thus informing all nodes in its range that they should refrain from transmission to prevent collision at the receiver. This solution also found application in cellular networks, in the *Cellular Digital Packet Data* (CDPD) standard, which provides an overlay packet mode access over circuit mode cellular networks [13].

A problem with BTMA and its variants is that a separate frequency band is needed for the busy tone. Radio propagation characteristics, however, depend on frequency, so the range for data and for the busy tone may not coincide, causing problems with protocols using multiple channels. This problem is solved by the *Multiple Access Collision Avoidance* (MACA) protocol [17] with a handshake of control packets. In MACA, the sender node A first sends a *request-to-send* (RTS) control packet

addressed to the intended receiver B, which replies with a *clear-to-send* (CTS) packet. On receiving the CTS, node A sends the data packet. If another node, C, also wanted to send data to B, then C, hearing the CTS of B, will know that B is busy, so it can refrain from transmission for the duration of the packet, which is included in the RTS control packet and copied into the CTS. If there is a collision of RTS packets at a destination, both senders use binary exponential backoff.

After a number of protocols applied the RTS/CTS concept, this line of development culminated in the IEEE 802.11 Wireless LAN standard [8]. In particular, the *distributed coordination function* (DCF) of IEEE 802.11 is a CSMA/CA access method utilizing several different IFSs. The standard also incorporates an optional access method called a *point coordination function* (PCF), which is essentially a centralized polling scheme. The DCF and PCF can coexist by having the two methods alternate, with a contention-free period followed by a contention period. HIPERLAN (*hi*gh-*per*formance *r*adio *l*ocal-*a*rea *n*etwork) is a set of wireless LAN standards used primarily in Europe [18]. There are two specifications (HIPERLAN1 and HIPERLAN2), which provide the features and capabilities similar to those of IEEE 802.11.

5.2. Combinations of Allocation and Contention

There are a number of other creative ways of combining the advantages of allocation and contention. The *Time-Spread Multiple Access* (TSMA) protocol [19] uses a fixed slot assignment, as in TDMA, but each node is assigned several slots in a frame. These slots are chosen by means of an algebraic method (finite, or Galois, fields), such that even if some of the transmissions may collide, eventually each node can successfully send a message in each frame; that is, collisions are resolved in a deterministic way, even though the frame length is much shorter than in TDMA. In an ad hoc network TSMA can provide deterministically bounded delay, which does not require rescheduling even if the network topology changes. It assumes, however, that the maximum nodal degree (number of neighbors of a node) remains bounded. This constraint is relaxed with the further idea of *protocol threading* [20] that interleaves several transmission schedules with different parameters.

Another combination is the *ADAPT* protocol [21], which utilizes the fact that in a TDMA schedule not all assigned slots are actually used for sending a packet, due to the random (bursty) nature of traffic. If a node leaves its assigned slot unused, then other nodes can sense this and can contend for the slot using an RTS/CTS based contention protocol. The CATA protocol [22] also incorporates contention within a slot, and provides explicit support for unicast, multicast, and broadcast packet transmissions. However, it is subject to instability at high traffic load due to the lack of a fixed frame length.

A very different type of combination is implemented in the *Meta-MAC* protocol [23]. Here a master protocol combines the "advice" of any set of MAC protocols that run independently. The combination is based on a weighted-majority decision with randomized rounding, using continuously updated weights depending on the feedback obtained from the channel. The Meta-MAC protocol can automatically and adaptively select the best protocol for the unknown or unpredictable conditions from a given set of protocols. While this set may contain different MAC protocols, it may instead consist of a single MAC protocol with different parameter settings. In this way Meta-MAC can be used for adaptively optimizing the parameters of a given MAC protocol for the current network conditions.

6. OUTLOOK

One might be tempted to think that the intensive research and development of MAC protocols, going on for several decades, has already produced all essential ideas in this field, and, consequently, that substantial new development is unlikely. This is, however, a wrong perception of this area. The design of MAC protocols critically depends on the technological constraints (examples of which are mentioned in Section 1). Thus, the emergence of new technologies and novel systems constantly poses new challenges to MAC protocol design. Let us mention a few examples of such emerging challenges.

Directional antennas can be used for radios, redefining the meaning of conflicts in packet radio networks [24] with the potential to improve spatial reuse. In *sensor networks* the sensor battery is a critical resource, as its replacement is seldom feasible. Even in more powerful wireless nodes, in addition to energy-efficient advances in hardware, corresponding improvements in software (i.e., protocols) to conserve energy at all layers of the protocol stack are required. In particular, the design of *energy-efficient* MAC protocols seek to reduce the energy wasted by nodes overhearing transmissions not intended for them [25]. Another interesting opportunity is to utilize the technological feasibility of more sophisticated radio hardware that can relax some of the traditional technological constraints. For example, it may be possible for a radio to implement a *multiple reception capability*, where more than one packet can be received successfully in the same time slot, utilizing the fact that some of the base technologies, such as CDMA, make it feasible [26]. In addition, new multimedia applications and services present many new challenges and opportunities for medium access control.

BIOGRAPHIES

Dr. Andras Farago received a B.S. in 1976, an M.S. in 1979, and a Ph.D. in 1981, all in electrical engineering from the Technical University of Budapest, Hungary. After graduation he joined the Department of Mathematics at the Technical University of Budapest. In 1982, he moved to the Department of Telecommunications and Telematics of the same university. He was also cofounder and research director of the High Speed Networks Laboratory, the first research center in high speed networking in Hungary. In 1997, he became Szechenyi Professor of Telecommunications at the Technical University of Budapest. In 1998, he joined the University of Texas at Dallas as a professor of Computer Science. His main

research area is in algorithms, protocols, and modeling of telecommunication networks. Dr. Farago authored over 100 research papers and in 1996 received the distinguished recognition Doctor of the Hungarian Academy of Sciences.

Violet R. Syrotiuk received her B.Sc. in 1983 from the University of Alberta, Canada, her M.Sc. in 1984 from the University of British Columbia, Canada, and her Ph.D. in computer science in 1992 from the University of Waterloo, Ontario, Canada. Dr. Syrotiuk is currently an assistant professor in the Department of Computer Science in the Erik Jonsson School of Engineering and Computer Science at the University of Texas at Dallas, where she is the codirector of the Scalable Network Engineering Techniques Laboratory (NET Lab). Dr. Syrotiuk's research has been funded by the Defense Advanced Research Projects Agency (DARPA), and is currently supported by grants from the National Science Foundation (NSF) and Raytheon Company. Her current research interests include medium access control (MAC) protocols with special emphasis on intelligent protocol adaptation to unknown or changing network conditions, and network layer protocols with an emphasis on scalable design.

BIBLIOGRAPHY

1. M. Ranhema, Overview of the GSM system and protocol architecture, *IEEE Commun. Mag.* 92–100 (April 1993).
2. D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, 1992.
3. I. Chlamtac, A. Faragó, and H. Zhang, A fundamental relationship between fairness and optimum throughput in TDMA protocols, *IEEE Int. Conf. Universal Personal Communications* (*ICUPC'96*), Cambridge, MA, Sept. 1996, pp. 671–675.
4. I. Chlamtac and S. Pinter, Distributed node organization algorithm for channel access in a multi-hop packet radio network, *IEEE Trans. Comput.* **36**(6): (1987).
5. C. Zhu and S. Corson, A five-phase reservation protocol (FPRP) for mobile ad hoc networks, *Proc. IEEE INFOCOM'98*, 1998.
6. A. Sen and M. L. Huson, A new model for scheduling packet radio networks, *IEEE INFOCOM'96*, 1996, pp. 1116–1124.
7. W. Stallings, *Wireless Communications and Networks*, Prentice-Hall, 2002.
8. Local and Metropolitan Area Networks Drafts (LAN/MAN 802), IEEE Standards Association Home Page, *http://standards.ieee.org*.
9. N. Abramson, The ALOHA system — another alternative for computer communications, *Proc. Fall Joint Computer Conf.*, 1970.
10. D. Aldous, Ultimate stability of exponential backoff protocol for acknowledgement based transmission control of random access communication channels, *IEEE Trans. Inform. Theory* **33**(2): 219–223 (1987).
11. F. P. Kelly, Stochastic models of computer communication systems, *J. Roy. Stat. Soc. B* **47**: 379–395 (1985).
12. J. Håstad, F. T. Leighton, and B. Rogoff, Analysis of backoff protocols for multiple access channels, *ACM Symp. Theory of Computing* (*STOC'87*), New York, May 1987, pp. 241–253.
13. S. Keshav, *An Engineering Approach to Computer Networking*, Addison-Wesley, 1997.
14. W. Crowther et al., A system for broadcast communication: Reservation-ALOHA, *Proc. 6th Hawaii Int. System Science Conf.*, Jan. 1973, pp. 596–603.
15. W. Stallings, *Data and Computer Communications and Networks*, 2nd ed., Macmillan, 1988.
16. A. Tobagi and L. Kleinrock, Packet switching in radio channels, Part II: The hidden terminal problem in carrier sense multiple access and the busy-tone solution, *IEEE Trans. Commun.* **23**: 1517–1453 (1975).
17. P. Karn, MACA — a new channel access protocol for packet radio, *ARRL/CRRL Amateur Radio 9th Computer Networking Conf.*, 1990, pp. 134–140.
18. European Telecommunications Standards Institute (ETSI), *http://www.etsi.org*.
19. I. Chlamtac and A. Faragó, Making transmission schedules immune to topology changes in multi-hop packet radio networks, *IEEE/ACM Trans. Network.* **2**(1): 23–29 (1994).
20. I. Chlamtac, A. Faragó, and H. Zhang, Time spread multiple access (TSMA) protocols for multihop mobile radio networks, *IEEE/ACM Trans. Network.* **5**(6): 804–812 (1997).
21. I. Chlamtac et al., ADAPT to mobility, *IEEE GLOBECOM*, Rio de Janeiro, Brazil, Dec. 1999.
22. Z. Tang and J. J. Garcia-Luna-Aceves, A protocol for topology-dependent transmission scheduling in wireless networks, *Proc. IEEE WCNC'99*, New Orleans, Sept. 21–24, 1999.
23. A. Faragó, A. D. Myers, V. R. Syrotiuk, and G. Záruba, Meta-MAC protocols: Automatic combination of MAC protocols to optimize performance for unknown conditions, *IEEE J. Select. Areas Commun.* **18**(9): 1670–1681 (2000).
24. Y.-B. Ko, V. Shankarkumar, and N. Vaidya, Medium access control protocols using directional antennas in ad hoc networks, *IEEE INFOCOM 2000*.
25. J. P. Monks, V. Bharghavan, and W.-M. W. Hwu, A power controlled multiple access protocol for wireless packet networks, *IEEE INFOCOM 2001*.
26. I. Chlamtac and A. Faragó, An optimal channel access protocol with multiple reception capacity, *IEEE Trans. Comput.* **43**(4): 480–484 (1994).

MEMS FOR RF/WIRELESS APPLICATIONS

HÉCTOR J. DE LOS SANTOS
Coventor, Inc.
Irvine, California

1. INTRODUCTION

The number of radiofrequency (RF) wireless applications and appliances is expected to explode in the first decade of the twenty-first century because of the unabated consumer demand for ubiquitous access to information [1]. The diversity of these consumers, which include both individuals and businesses, and the nature of the information they demand, which encompasses not only voice communications but also video, broadband data, messaging, navigation, direct broadcast satellite links, and the Internet, in the context of global connectivity, imposes, in turn,

such extreme levels of functionality and sophistication on these appliances, that doubts have been cast on the ability of conventional integrated circuit (IC) technology and fabrication techniques to deliver the high-performance RF functions required [1–3]. The seriousness of the matter may be gauged from an examination of the evolution in wireless standards (Table 1). In particular, while the first-generation (1G) appliances provided only single-band analog cellular connectivity capabilities, those of the second (2G) had to provide dual-mode dual-band digital voice plus data, and now those of the third (3G) and fourth (4G) generations will have to provide multimode (i.e., analog/digital), multiband (i.e., various frequencies), and multistandard [i.e., various standards — Global System for Mobile Communications (GSM) — a leading digital cellular system, which allows eight simultaneous calls on the same radiofrequency, digital European Cordless Telecommunications (DECT) — a system for the transmission of integrated voice and data in the range of 1.8–1.9 GHz, cellular digital packet data (CDPD) — a data transmission technology that uses unused cellular channels to transmit data in packets in the range of 800–900 MHz, General Packet Radio Service (GPRS) — a standard for wireless communications that runs at 150 kilobits per second (kbps), and code-division multiple access (CDMA) — a North American standard for wireless communications that uses spread-spectrum technology to encode each channel with a pseudorandom digital sequence] performance capabilities. Furthermore, the desire to maintain seamless connectivity, on a global basis, as the user moves through independently operated Internet Protocol (IP) networks [4–5], such as among various countries, dictates that these appliances be equipped for operation over a wide variety of access and network technologies and standards, with their accompanying processing overhead associated with function management. This latter requirement, in view of limited battery power, makes power consumption minimization a prime factor in the successful implementation of these systems.

Beyond technical performance considerations, however, commercial success is largely dependent on achieving a cost-effective solution. Thus, as the need for interfacing with off-chip components reflects adversely on the manufacturer's ability to meet cost constraints, its avoidance provides a strong motivation for the exploration of alternatives. Potential paths toward these alternatives become readily manifest when the limitations of conventional IC technologies for implementing RF functions are examined. In particular, attempts produce high-quality on-chip passive RF components, such as inductors, capacitors, varactors, switches, resonators, and transmission lines [2] — the core of wireless functions — reveal that this is a virtually impossible task because of the poor properties of silicon substrates. Against this bleak picture, microelectromechanical systems (MEMS) technology is emerging as the disruptive technology whose versatile fabrication techniques might well provide the solution to these limitations, insofar as it is poised to render virtually parasitic-free passive RF devices, side by side with the electronics, while simultaneously reaping the low cost property that characterizes batch fabrication processes.

In this article we review the fundamentals of MEMS fabrication techniques, the performance of typical RF devices exploiting them, and the RF MEMS circuits and systems it enables.

2. FUNDAMENTALS OF MEMS FABRICATION TECHNIQUES

MEMS fabrication techniques enable the construction of three-dimensional mechanical structures in the context of the conventional process utilized in the production of planar integrated circuits. As such, the structures created encompass feature sizes between a few and several hundred micrometers. To introduce the various approaches to MEMS fabrication, we begin with a brief review of the conventional IC fabrication process [2].

Table 1. Wireless Standards: The Evolution Blueprint

1G	2G	3G	4G
Analog cellular (single band)	Digital (dual-mode, dual band)	Multimode, multiband software-defined radio	Multistandard + multiband
Voice telecom only	Voice + data telecom	New services markets beyond traditional telecom: *higher-speed data, improved voice, multimedia mobility*	
Macrocell only	Macro/micro/picocell	Data networks, Internet, VPN, *WINternet*	
Outdoor coverage	Seamless indoor/outdoor coverage		
Distinct from PSTN	Complementary to fixed PSTN		
Business customer focus	Business + consumer	Total communications subscriber: virtual personal networking	

Source: http://www .uwcc.org.

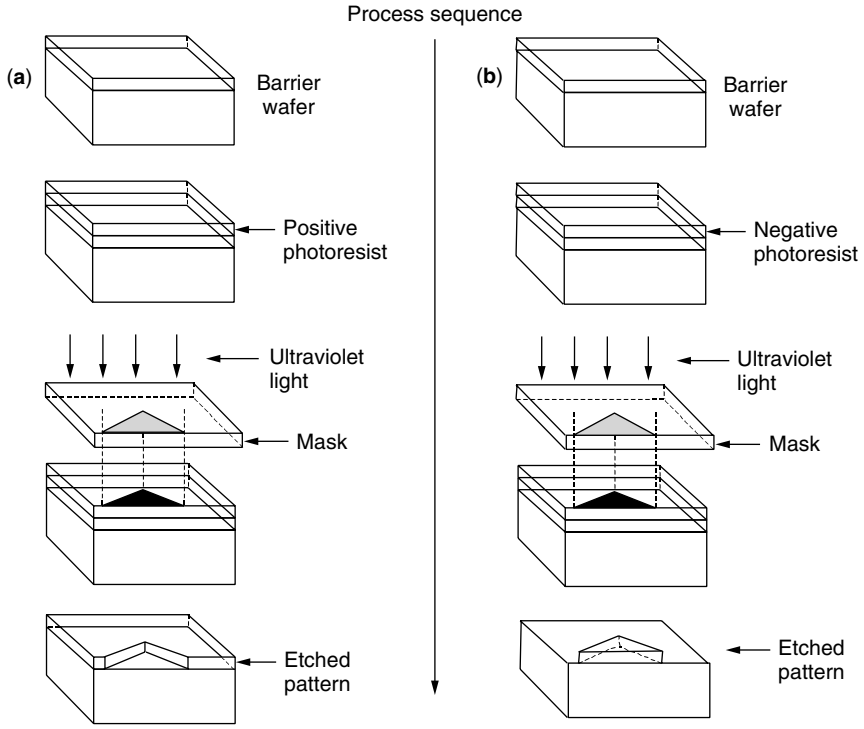


Figure 1. Conventional IC fabrication process using (**a**) positive and (**b**) negative photoresist.

2.1. Conventional IC Fabrication Process

The conventional IC fabrication process employs a sequence of photolithography and chemical etching steps to transfer a layout pattern onto the surface of a wafer. The process is illustrated by the sketch of Fig. 1. A semiconductor wafer, covered with a barrier material, is coated with a soft light-sensitive material called photoresist (PR). The PR may be positive or negative in the sense that, when exposed to ultraviolet (UV) light, it may harden or weaken, respectively. Thus, when positive PR is exposed through a mask containing transparent and opaque regions, representing the pattern to be transferred, a pattern identical to that on the mask is defined on the barrier upon subsequent chemical etching (Fig. 1a). On the other hand, if the PR is negative, the negative image of the pattern in the mask ends up being defined after etching (Fig. 1b).

2.2. RF MEMS Fabrication Approaches

Two main approaches, summarized below, dominate those employed to build three-dimensional mechanical structures for RF MEMS, namely, surface micromachining and bulk micromachining. Further information on these and other RF MEMS fabrication alternatives may be found in an earlier treatise [2].

2.2.1. Surface Micromachining. In surface micromachining, freestanding micromechanical structures are formed on the surface of a wafer by depositing and patterning a sequence of thin film material layers. Those layers that are deposited, and later removed, are called *sacrificial layers*; those layers that remain freestanding are called *structural layers*. Thus, the creation of every freestanding element involves the following main steps: (1) depositing a sacrificial layer; (2) opening a hole on the sacrificial layer that will permit access to the underlying wafer or structural layer; (3) deposition and patterning of the structural layer, such that it becomes anchored in the underlying substrate via a connection through the hole that was opened in the sacrificial layer; and (4) removal of the sacrificial layer to *release* the structure from it. These steps are sketched in Fig. 2. Examples of RF MEMS applications using this technique include switches, inductors, and varactors.

One fundamental limitation of surface micromachining [2] is the phenomenon of stiction. *Stiction* refers to the propensity of microscopic structures to stick together when they are close to each other or in a humid environment. An example of the former may occur when they experience van der Waals and electrostatic forces (due to random charging), whereas an example of the latter is when the structure is immersed in a wet etchant for dissolving the sacrificial layer, in which case the surface tension of the etchant may overcome the springback force that attempts to bring the structure to its equilibrium configuration. Popular approaches to overcome stiction during the release process involve the adoption of dry-etching chemistries, drying of the released wafer with supercritical CO_2, or freezing and then sublimating the release liquid.

2.2.2. Bulk Micromachining. In bulk micromachining, micromechanical structures are sculpted within the confines of a wafer. This is accomplished by the ingenious exploitation of highly directional (anisotropic) and nondirectional (isotropic) etchants, together with their etching rates, in relation to the various crystallographic planes of the wafer. Essentially, planes with a higher density of atoms will etch more slowly. Similarly, by defining heavily doped contour layers and pn (positive–negative) junctions,

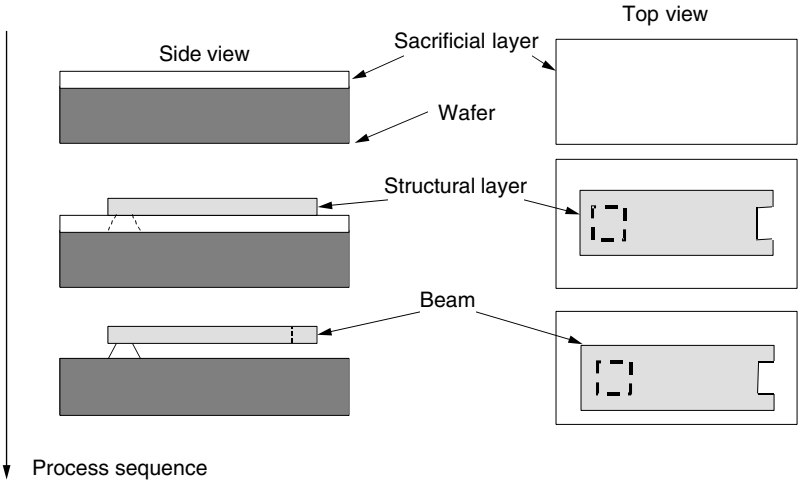


Figure 2. Sketch of surface micromachining process.

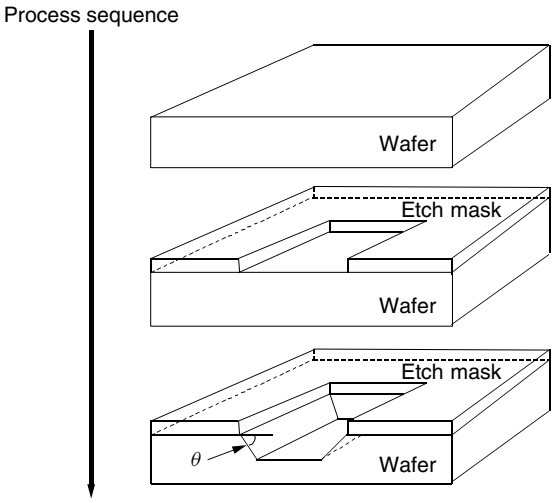


Figure 3. Sketch of bulk micromachining process.

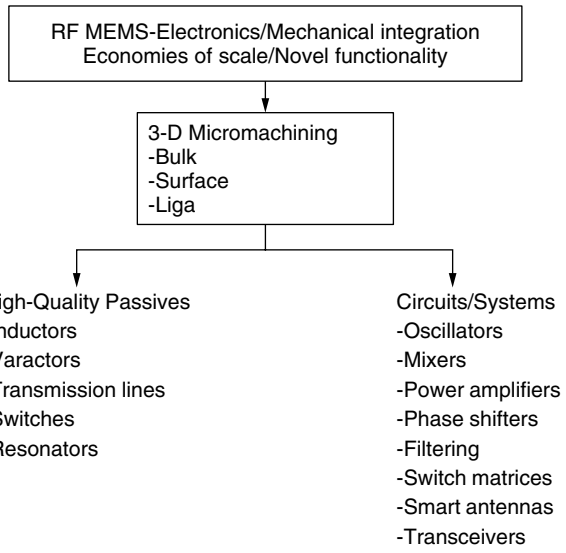


Figure 4. RF components enabled by MEMS technology (after Ref. 1).

for slowing down or totally stopping the etching process, respectively, the technique allows the creation of deep cavities. Figure 3 shows sketches of bulk micromachined structures. Examples of RF MEMS applications using this technique include transmission lines and inductors.

One fundamental limitation of bulk micromachining [2] is that the aspect ratio of the sculpted structures, such as the slope or verticality of the cavity walls, is a function of the angle between crystallographic planes. To overcome this limitation, a new technique called *deep reactive-ion etching* has been introduced.

3. RF MEMS DEVICES, CIRCUITS, AND SYSTEMS

The high level of interest in RF MEMS for wireless applications stems from the versatility of its fabrication approaches for producing virtually ideal (parasite-free) RF devices, in particular, in conjunction with integrated circuits, thus enabling new levels of circuits and systems functionality, together with the potential for achieving overall reductions in systems' weight/size and power, while exploiting the economies of scale germane to ICs. Figure 4 captures the RF MEMS arsenal and its potential areas of influence in wireless communications.

3.1. RF MEMS Passive Devices

Virtually all types of passive devices utilized in wireless applications (i.e., inductors, capacitors, varactors, transmission lines, switches, and resonators) have been demonstrated via MEMS fabrication techniques. In what follows, we describe representative examples of each of these.

3.1.1. Inductors. The performance of integrated inductors, in particular, their self-resonance frequency (the maximum frequency delimiting inductive behavior), and quality factor, is well known to be limited by the capacitance and resistance of the substrate on which they are disposed. Accordingly, a number of approaches aimed at separating the trace structure from the substrate have been advanced, including

1. Creating an airpit under the trace spiral via bulk micromachining [6], (Fig. 5a), which resulted

in a self-resonance frequency enhancement from 800 MHz to 3 GHz, on substrate removal, and a Q of 22 at 270 MHz on an 115-nH inductor;

2. Suspending the trace spiral a distance over the wafer surface via a combination of bulk and surface micromachining [8] (Fig. 5b), which resulted in a Q of 30 at 8 GHz on a 10.4-nH inductor with a self-resonance frequency of 10.1 GHz;
3. Implementing the inductor as a solenoid via surface micromachining [9] (Fig. 5c), which resulted in a Q of 16.7 at 2.4 GHz on 2.67-nH inductors;
4. Using self-assembly techniques to erect the plane of trace structure perpendicular to the substrate [10] (Fig. 5d), which resulted in improvements in the Q of 2nH meander inductors from 4 at 1 GHz, for the planar realization, to 20 at 3 GHz for the self-assembly implementation

3.1.2. Varactors. Varactors are indispensable in the operation of voltage-controlled oscillators (VCOs). High-quality varactors, however, are difficult to produce in the context on an IC because of processes are usually optimized for other devices, such as transistors [2]. Since MEMS devices may be integrated on chip without disrupting the process flow, specifically in a postprocessing step, several schemes using surface micromachining have been exploited to create potentially IC-compatible varactors. These are predicated upon varying one of the parameters defining capacitance, $C = \varepsilon A/d$, where ε is the dielectric constant, A is the area, and d is the plate separation. Schemes that vary the plate separation employ a square plate suspended by four beams and disposed over a bottom plate (electrode). At zero bias, there is a maximum distance between top and bottom plates. However, when a voltage is applied between the plates, the force of electrostatic attraction between them causes the gap to diminish [2], thus varying (increasing) the capacitance [11]. In another scheme, the capacitor structure consists of interdigitated plates. Thus by varying the degree of engagement, namely the effective device area, the overall capacitance is made to vary [12]. Finally, in a more recent scheme (Fig. 6), the effective dielectric constant of the structure is made to vary by sliding, in a lateral fashion, a dielectric between the parallel plates of the capacitor [13].

3.1.3. Transmission Lines. The performance of transmission lines, in particular, their attenuation, is a strong

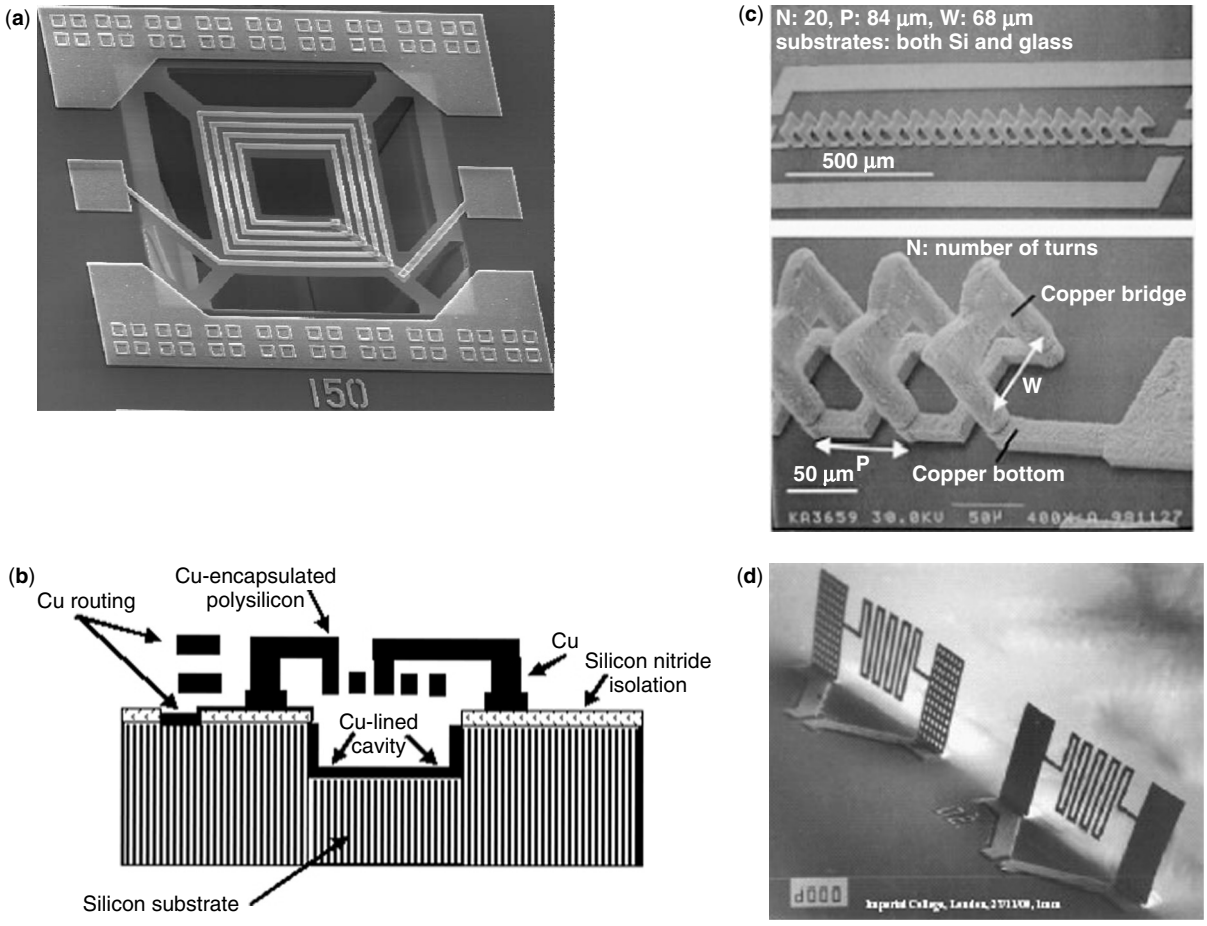


Figure 5. RF MEMS Inductors: (**a**) bulk-micromachined inductor [7]; (**b**) schematic of a copper-encapsulated polysilicon inductor suspended over a copper-lined cavity beneath [8]; (**c**) SEM photograph of 20-turn, on-Si, air-core, all-copper solenoid inductor (*upper* — overview; *lower* — magnified view) [9]; (**d**) $4\frac{1}{2}$-turn meander inductor after self-assembly [10].

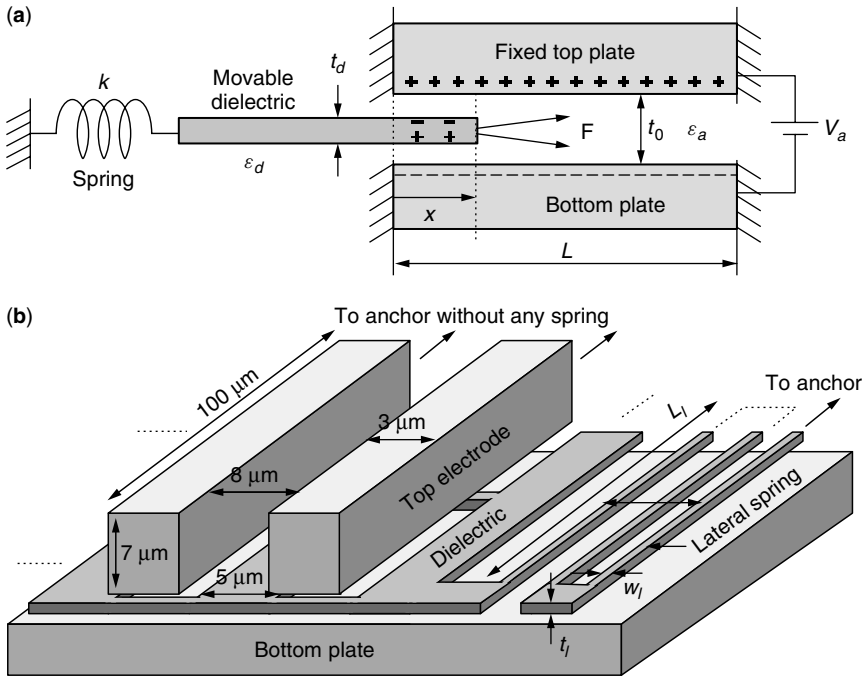


Figure 6. Micromachined varactor: (**a**) conceptual schematic; (**b**) actual implementation using a lateral spring. (from Ref. 13 © 1998 IEEE).

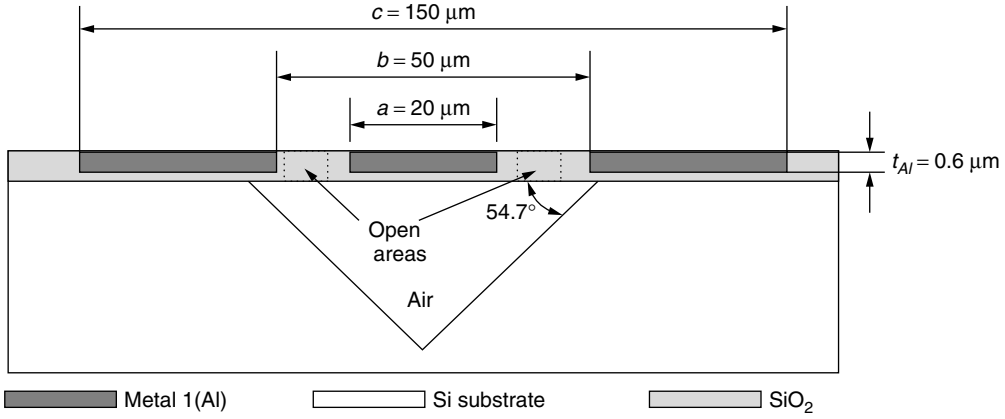


Figure 7. Cross-sectional view of bulk-etched transmission-line structure (after Ref. 14).

function of the substrate that mechanically supports them. Thus, silicon transmission lines tend to be lossy. To enable low-loss interconnects in the context of a silicon IC, Milanovic et al. [14] developed the structure shown in Fig. 7. This is a coplanar waveguide (CPW) transmission line, in which, by opening access windows in the top passivation, an airpit is formed underneath the center conductor to eliminate the substrate under it and, thus, minimize the structure's insertion loss (IL). Improvements in the IL of about 7 dB at 7 GHz, and 20 dB at 20 GHz, with respect to the nonetched reference, were obtained.

3.1.4. Switches. Switches may be considered one of the RF MEMS elements with the potential for greatest impact in wireless communications. With such capabilities as [3] series resistance $<1\ \Omega$, insertion loss at 1 GHz within 0.1 dB, Isolation at 1 GHz $>$ 40 dB, IP3 $>$ 66 dBm, 1 dB compression $>$33 dBm, size $<1\ mm^2$, switching speed of the order of 1 μs, control voltage between 3 and 30 V, and control current $<1\ \mu A$, with no standby power consumption, they have become the potential enabler for many systems, in particular, phased arrays and switch matrices [2]. Figure 8 [15] shows the structure and operation of a state-of-the-art RF MEMS switch. The switch consists of a metal bridge bar that is moved to make or break contact with an underlying signal line. Bridge motion is achieved by voltage biasing a mechanical actuator supporting the bridge. The device covers an area of approximately $250 \times 250\ \mu m$. The typical performance included and effective capacitance of 2fF in the OFF state, for an isolation of 30 dB at 40 GHz, an effective resistance of 1 Ω, giving an insertion loss of 0.2 dB, and a return loss of 25 dB. In addition, the actuation voltage was 85 V and the switching time about 10 μs.

3.1.5. Resonators. Resonators are essential elements for realizing filters and oscillators [2]. Their RF MEMS implementations take usually two main forms: a *cavity*,

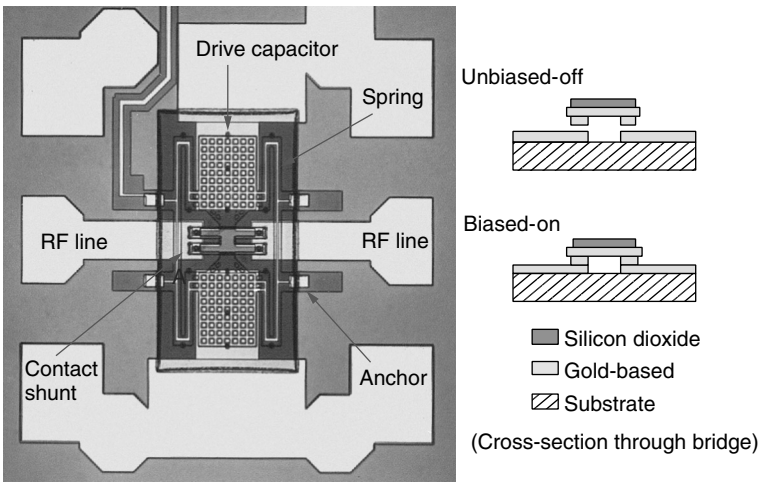


Figure 8. Structure of RF MEM switch (courtesy of Drs. R. E. Mihailovich and J. DeNatale, Rockwell Scientific) (from Ref. 15 © 2001 IEEE).

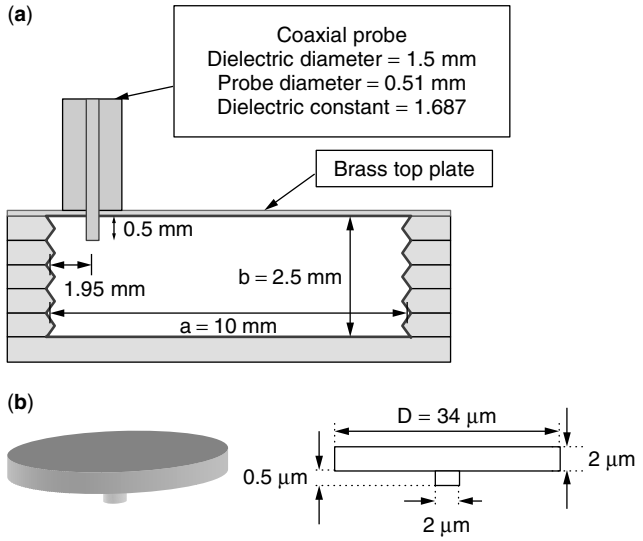


Figure 9. RF MEMS resonators: (**a**) schematic and photograph of bulk micromachined cavity resonator at 30 GHz (length $c = 5$ mm) (courtesy of Mr. M. Stickel and Prof. G. V. Eleftheriades, Univ. of Toronto); (**b**) schematic of contour-mode disk resonator (courtesy of Mr. Hideyuki Maekoba, Coventor, Inc.).

for applications beyond 10 GHz, and a *micromechanical resonator*, for applications below 1 GHz. Figure 9a shows an example of the former, which exhibited $Q > 2000$ at 30 GHz, and Fig. 9b, which exhibited $Q = 9200$ at 156 MHz, an example of the latter.

In addition to these resonators, there is the film bulk acoustic wave resonator (FBAR), which is based on the formation of an acoustic cavity out of a piezoelectric material, and which exhibits Q between 500 and >1000, at frequencies of several GHz [16].

3.2. Circuit Applications of RF MEMS Devices

While a number of RF MEMS-based circuits, notably, oscillators and filters, have been demonstrated [16–21], phase shifters exploiting MEM switches may be considered the major technology driver, as they are an enabling component for the realization of large phased arrays. An example of such a phase shifter is shown in Fig. 10 [23]. This is a line-switched true(real)-time delay (TTD) 4-bit phase shifter implemented with the RF MEMS switches described in Section 3.1.4. In the DC 40-GHz frequency band the circuit exhibited delay times in the 106.9–193.9 ps range. This was accomplished with a resolution of 5.8-ps-delay increments, which represents a phase shift of 22.5° at 10.8 GHz produced by using microstrip lines with a length of 600 μm. The total chip area was 6×5 mm^2.

4. SUMMARY

In this article we have presented a brief review of MEMS for RF/wireless applications. In particular, we have addressed the motivations propelling the high level of interest in this emerging technology, its fabrication fundamentals, and the sample realizations and performance of key devices that it enables, namely,

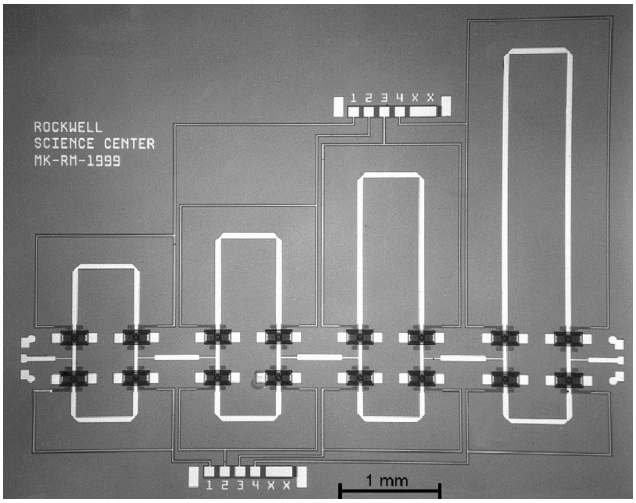


Figure 10. Photograph of 4-bit RF MEMS TTD phase shifter. The second longest bit (bit 3) was fabricated separately to analyze both the insertion loss and the impact of the matching section on TTD performance. (*Source*: Ref. 23 © 2001 IEEE. Courtesy of Drs. R. E. Mihailovich and J. DeNatale.)

inductors, varactors, transmission lines, switches, and resonators. We have also presented, perhaps, the major RF MEMS technology driver, namely, the phase shifter circuit, which is of great importance to large systems applications, in particular, phased arrays.

BIOGRAPHY

Héctor J. De Los Santos is Principal Scientist at Conventor, Inc., Irvine, California, where he leads Conventor's RF MEMS R&D. He received a Ph.D. from the School of Electrical Engineering, Purdue University, West Lafayette, Indiana, in 1989. From March 1989 to September 2000, he was employed at Hughes space and Communications Company, Los Angeles, where he served as Scientist and Principal Investigator and Director of the Future Enabling Technologies IR&D Program. Under this program he pursued research in the areas of RF MEMS, quantum functional devices and circuits, and photonic bandgap devices and circuits. Dr. De Los Santos holds a dozen patents and has over six patents pending. He is author of the bestseller textbook *Introduction to Microelectromechanical (MEM) Microwave Systems*, Artech House, Norwood, Massachusetts, 1999, and of the book *RF MEMS Circuit Design for Wireless Communications*, Artech House, June 2002. Dr. De Los Santos is a Senior Member of the IEEE, and member of Tau Beta Pi, Eta Kappa Nu, and Sigma Xi. He is an IEEE Distinguished Lecturer of the Microwave Theory and Techniques Society for the 2001–2003 term.

BIBLIOGRAPHY

1. H. J. De Los Santos, MEMS — a wireless vision, *Proc. 2001 International MEMS Workshop*, Singapore, July 4–6, 2001.
2. H. J. De Los Santos, *Introduction to Microelectromechanical (MEM) Microwave Systems*, Artech House, Norwood, MA, 1999.
3. R. J. Richards and H. J. De Los Santos, MEMS for RF/wireless applications: The next wave, *Microwave J.* (March 2001).
4. R. R. Parrish, Mobility and the Internet, *IEEE Potentials Mag.* 8–10 (April/May 1998).
5. A. Fasbender, F. Reichert, E. Geulen, and J. Hjelm, Any network, any terminal, anywhere, *IEEE Pers. Commun. Mag.* 22–30 (April 1999).
6. J.-Y. Chang, A. A. Abidi, and M. Gaitan, Large suspended inductors on silicon and their use in a 2um CMOS RF amplifier, *IEEE Electron Device Lett.* **14**: 246–248 (1993).
7. Y. Sun, H. van Zeijl, J. L. Tauritz, and R. G. F. Baets, Suspended membrane inductors and capacitors for application in silicon MMICs, *IEEE Microwave and Millimeter-wave Monolithic Circuits Symp. Digest of Papers*, 1996, pp. 99–102.
8. H. Jiang, Y. Wang, J.-L. A. Yeh, and N. C. Tien, Fabrication of high-performance on-chip suspended spiral inductors by micromachining and electroless copper plating, *2000 IEEE IMS Digest of Papers*, Boston, MA.
9. J.-B Yoon et al., Surface micromachined solenoid On-Si and On-Glass inductors for RF applications, *IEEE Electron Device Lett.* **20**: 487 (1999).
10. G. W. Dahlmann et al., MEMS high Q microwave inductors using solder surface tension self-assembly, *2001 IEEE IMS Digest of Papers*.
11. D. J. Young and B. E. Boser, A micromachined variable capacitor for monolithic low-noise VCOs, *Hilton Head '96*, pp. 86–89.
12. J. J. Yao, Topical review: RF MEMS from a device perspective, *J. Micromech. Microeng.* **10**: R9–R38 (2000).
13. J.-B. Yoon and C. T.-C. Nguyen, A high-Q tunable micromechanical capacitor with movable dielectric for RF applications, *1998 IEEE Int. Electron Devices Meeting Digest of Papers*, pp. 489–492 (Figs. 1, 4, 6).
14. V. Milanovic et al., Micromachined microwave transmission lines in CMOS technology, *IEEE Trans. Microwave Theory Tech.* **45**: 630–635 (1997).
15. R. E. Mihailovich et al., MEM relay for reconfigurable RF circuits, *IEEE Microwave Wireless Components Lett.* **11**: 53–55 (Feb. 2001).
16. H. J. De Los Santos, *RF MEMS Circuit Design for Wireless Communications*, Artech House, Norwood, MA, 2002.
17. P. Bradley, R. Ruby, and J. D. Larson III, A film bulk acoustic resonator (FBAR) duplexer for USPCS, *2001 IEEE Int. Microwave Symp.*, Phoenix, AZ.
18. A. R. Brown and G. M. Rebeiz, A high-performance integrated-band diplexer, *IEEE Trans. Microwave Theory Tech.* **47**: 1477–1481 (Aug. 1999).
19. H.-T. Kim, J.-H. Park, Y. Kim, and Y. Kwon, Millimeter-wave micromachined tunable filters, *1999 IEEE MTT-S Digest*, pp. 1235–1238.
20. F. D. Bannon III, J. R. Clark, and C. T.-C. Nguyen, High-Q HF microelectromechanical filters, *IEEE J. Solid-State Circuits* **35**: 512–526 (April 2000).
21. K. Wang and C. T.-C. Nguyen, High-order micromechanical electronic filters, *Proc. IEEE Micro Electro Mechanical Systems Workshop*, 1999, pp. 25–30.
22. C. T.-C. Nguyen and R. T. Howe, An integrated CMOS micromechanical resonator high-Q oscillator, *IEEE J. Solid-State Circuits* **34**: 440–445 (April 1999).
23. M. Kim, J. B. Hacker, R. E. Mihailovich, and J. F. DeNatale, A DC-40 GHz four-bit RF MEMS true-time delay network, *IEEE Microwave Wireless Components Lett.* **11**: 56–58 (Feb. 2001).

MICROSTRIP ANTENNAS

NAFTALI HERSCOVICI
Anteg, Inc.
Framingham, Massachusetts

1. INTRODUCTION

Microstrip antennas consist of a patch of metalization separated from a ground plane by a dielectric substrate (Fig. 1). The concept of the microstrip radiator was proposed in the early 1950s by Deschamps [1]. Only a couple of years after Deschamps' communication, a French patent on a similar geometry was awarded to Gutton and Baissinot [2]. However, no reports on this subject were published in the literature until the early

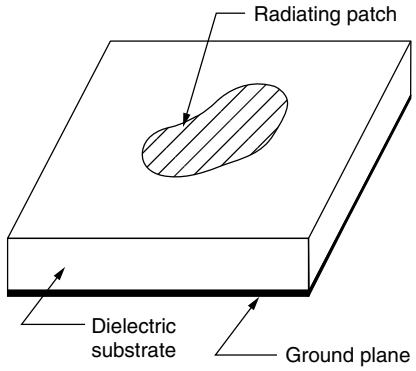


Figure 1. The microstrip antenna.

1970s, when Byron [3] proposed the "conductive strip radiator separated from the ground plane by a dielectric substrate." Since then, to the present, significant progress has been made in the development of dielectric substrates and computer technologies, which both contributed to the development of numerous variations of the basic concept proposed by Deschamps.

Since microstrip radiators are essentially planar in nature [two-and-one-half-dimensional ($2\frac{1}{2}$D) structures], they are narrowband antennas. This is true for the initial configurations, which consisted of a single patch. Since the late 1980s, considerable effort has been invested in developing broadband microstrip elements, which were succeeding in approaching 90% bandwidth for a VSWR < 2.

Microstrip antennas have a number of advantages in comparison with other types of antennas:

- Light weight, low volume, and, to a certain extent, flexibility, which allows integration on conformal surfaces
- Allow integration with active devices
- Easily arrayable, allowing a significant freedom of design and synthesis of various radiation patterns.
- Low fabrication cost
- All polarizations possible with relatively simple feeding mechanisms

The limitations of microstrip antennas and arrays are

- Narrow band.
- Large arrays can exhibit low efficiency due to the losses associated with the feeding network.
- Radiation coverage limited to one hemisphere.
- Very low sidelobe arrays are difficult to obtain because of the radiation of the feeding network.
- Losses associated with surface waves.
- Low power handling capability.

These limitations of basic microstrip antennas and arrays can be overcome with more sophisticated architectures, which might make the design expensive to mass production and sensitive to manufacturing tolerances.

In spite of their simple geometry, the design of microstrip antennas can be a complicated and iterative process. This is mostly because of their high-Q nature and the complexity of the analysis associated with the accurate modeling such structures. Many approximate models have been developed and, since the late 1980s, with the development of fast computers, numerical methods have been developed for accurate analysis.

The approximate methods treat the microstrip patch antenna as a transmission line (the *transmission-line model* and its derivatives) or a cavity (the *cavity model*) and provide a better physical insight, which is missing in the accurate CAD models. The formulas for the main characteristics of the microstrip antennas given below are based on approximate methods that are accurate enough for the initial design iteration.

2. ELECTRICAL CHARACTERISTICS OF A RECTANGULAR PATCH ANTENNA

2.1. Radiation Characteristics of the Rectangular Microstrip Antenna Element

Considering its Cartesian shape, the rectangular microstrip patch antenna is relatively easy to analyze, so the electrical characteristics of microstrip antennas presented below pertain to the rectangular patch. Furthermore, experience shows that the rectangular patch is much more used than any other type of patch. The typical rectangular microstrip radiator is shown in Fig. 2. For calculation of the radiation patterns, the rectangular patch can be seen as a line resonator, approximately a half-wavelength long [4]. The radiation occurs mostly from the fringing fields at the open-transmission-line ends (Figs. 3 and 4). Including the effect of the ground plane and substrate, the E-plane pattern is given by

$$E_\theta(\theta) = -jk_0V_0W\frac{e^{-jk_0r}}{2\pi r}\sin c\left(\frac{k_0h}{2}\sin\theta\right) \times \cos\left(\frac{k_0L}{2}\sin\theta\right)F_1(\theta) \tag{1}$$

$$E_\phi = 0 \tag{2}$$

and the H-plane pattern is given by

$$E_\phi(\theta) = jk_0V_0W\frac{e^{-jk_0r}}{2\pi r}\sin c\left(\frac{k_0W}{2}\sin\theta\right)\cos\theta F_2(\theta) \tag{3}$$

$$E_\theta(\theta) = 0 \tag{4}$$

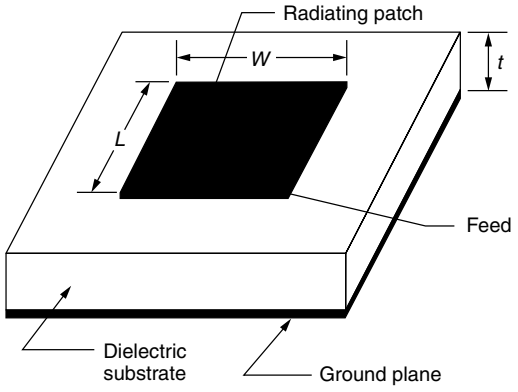


Figure 2. The rectangular, single-layer microstrip radiator.

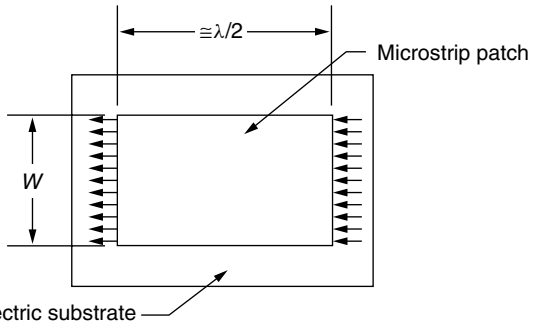


Figure 3. The rectangular microstrip patch with equivalent radiating slots.

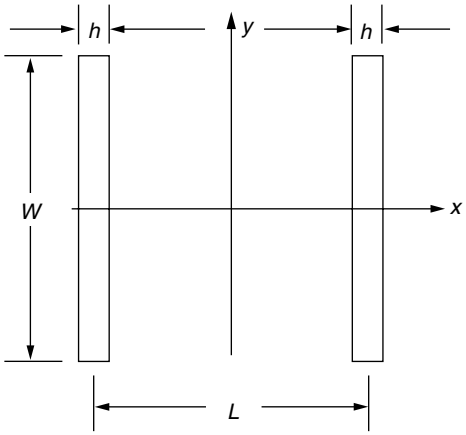


Figure 4. The rectangular microstrip antenna represented as two radiating slots.

where

$$F_1 = \frac{2\cos\theta\sqrt{\varepsilon_r - \sin^2\theta}}{\sqrt{\varepsilon_r - \sin^2\theta} - j\varepsilon_r\cos\theta\cot(k_0 h\sqrt{\varepsilon_r - \sin^2\theta})} \tag{5}$$

$$F_2 = \frac{2\cos\theta}{\cos\theta - j\sqrt{\varepsilon_r - \sin^2\theta}\cot(k_0 h\sqrt{\varepsilon_r - \sin^2\theta})} \tag{6}$$

and

W = the width of the patch (H-plane dimension)
L = the length of the patch (E-plane dimension also called the resonant dimension)
h = the thickness of the substrate
ε_r = the dielectric constant of the substrate
$k_0 = \dfrac{2\pi}{\lambda_0}$, where λ_0 is the wavelength

The half-power beamwidth can be approximated to give

$$\theta_{BH} = 2\arccos\sqrt{\frac{1}{2\left\{1 + \dfrac{k_0 W}{2}\right\}}} \tag{7}$$

$$\theta_{BE} = 2\arccos\sqrt{\frac{7.03}{3k_0^2 L^2 + k_0^2 h^2}} \tag{8}$$

2.2. Input Impedance of the Rectangular Microstrip Antenna Element

The transmission-line model [4] does not take the position of the feeding point along the length of the patch into consideration. Newman and Tulyathan proposed [5] an improved model, which solves this problem. They derived a simple formula for the input impedance, which also includes the reactance of the probe:

$$Z_{\text{in}} = Z_1 + jX_L \tag{9}$$

$$Z_1 = \frac{1}{Y_1} \tag{10}$$

$$Y_1 = Y_0\left[\frac{Z_0\cos\beta L_1 + jZ_w\sin\beta L_1}{Z_w\cos\beta L_1 + jZ_0\sin\beta L_1} + \frac{Z_0\cos\beta L_2 + jZ_w\sin\beta L_2}{Z_w\cos\beta L_2 + jZ_0\sin\beta L_2}\right] \tag{11}$$

where

$$X_L = \frac{377}{\sqrt{\varepsilon_r}}\tan\left(\frac{2\pi h}{\lambda_0}\right) \tag{12}$$

is the probe reactance, Z_0 is the characteristic impedance of the microstrip, and $Y_W = 1/Z_W$ is the wall admittance in the E plane as defined in the cavity model proposed by Bahl [6]. Then

$$Y_W = G_W + jB_W \tag{13}$$

$$G_W = \frac{0.00836W}{\lambda_0} \tag{14}$$

$$B_W = 0.01668\frac{\Delta I}{h}\frac{W}{\lambda_0}\varepsilon_e \tag{15}$$

where ε_e is the dielectric effective constant, given by

$$\varepsilon_e = \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 1}{2}\left(1 + \frac{12h}{W}\right)^{-1/2} \tag{16}$$

and Δl is a length correction factor given by

$$\Delta l = 0.412h\frac{(\varepsilon_e + 0.3)(W/h + 0.264)}{(\varepsilon_e - 0.258)(W/h + 0.8)} \tag{17}$$

Again, the quantities defined above are approximations required by the transmission-line model (and other approximate models) and for the same quantities, various expressions have been derived [6,7].

2.3. Design Procedure for Rectangular Microstrip Antennas

2.3.1. Choice of the Dielectric Substrate. The first step in the design of microstrip antenna is the choice of the dielectric substrate, which includes the following considerations: dielectric constant, thickness, and losses. As shown further, the Q factor of the antenna is strongly dependent on the dielectric constant and substrate thickness, as well as the efficiency associated with the surface wave excitation.

2.3.2. The Element Width. Once the dielectric substrate is chosen, the "effective dielectric constant" of the

substrate has to be calculated [Eq. (16)]. The dielectric constant ε_r has a definite impact on the resonance frequency which is determined not only by ε_r and h but also by the length of the patch, L. The width of the patch has an impact on the input impedance and a good approximation for W is

$$W = \frac{c}{2f_r}\sqrt{\frac{2}{\varepsilon_r + 1}}$$

where c is the velocity of light and f_r is the resonance frequency. The width of the patch, which also controls the radiation pattern [Eq. (7)], has to be chosen carefully to avoid the excitation of higher-order modes. The dependence of W on frequency for three different dielectric constants is shown in Fig. 5 [8].

2.3.3. The Element Length. The element length (often known as the *resonant dimension* of the rectangular patch) determines the resonant frequency. Here *resonant frequency* means "the frequency where the reactance of the antenna equals zero." In practice, the presence of additional components, such as feeding probes and stubs, alter the meaning of this definition, which is sometimes changed to "the frequency where the maximum of the real part of the input impedance occurs."

Knowing ε_e and Δ_l one can calculate L (the resonance dimension of the rectangular patch):

$$L = \frac{c}{2f_r\sqrt{\varepsilon_e}} - 2\Delta l \tag{18}$$

Figure 6 shows L versus the resonance frequency for a number of different substrates.

2.3.4. Radiation Patterns. Unlike other types of antennas where the characteristics of the radiation patterns are determined by parameters that do not significantly impact the input impedances and working bandwidth, in the case of microstrip elements, once the dielectric substrate, the patch length, and the patch width are determined, the characteristics of the radiation pattern are already set. To obtain narrower beamwidths (higher directivity) using microstrip radiators, arrays will have to be used.

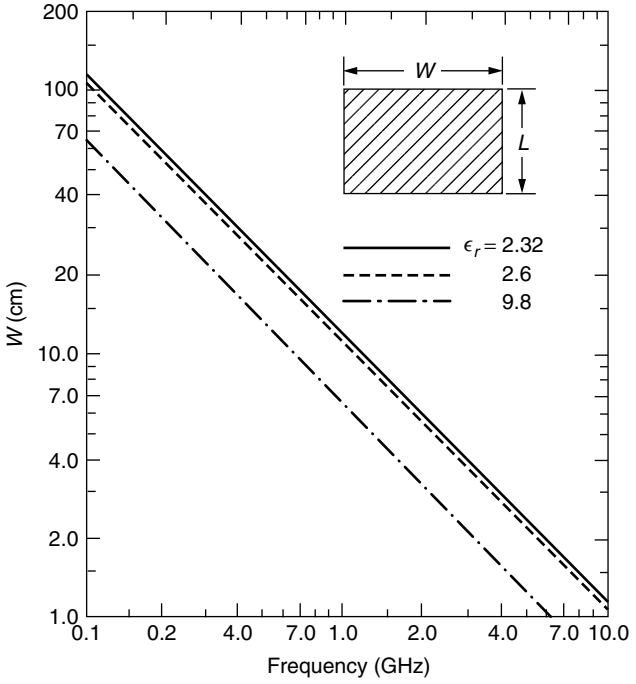


Figure 5. Element width versus frequency for different dielectric substrates [8].

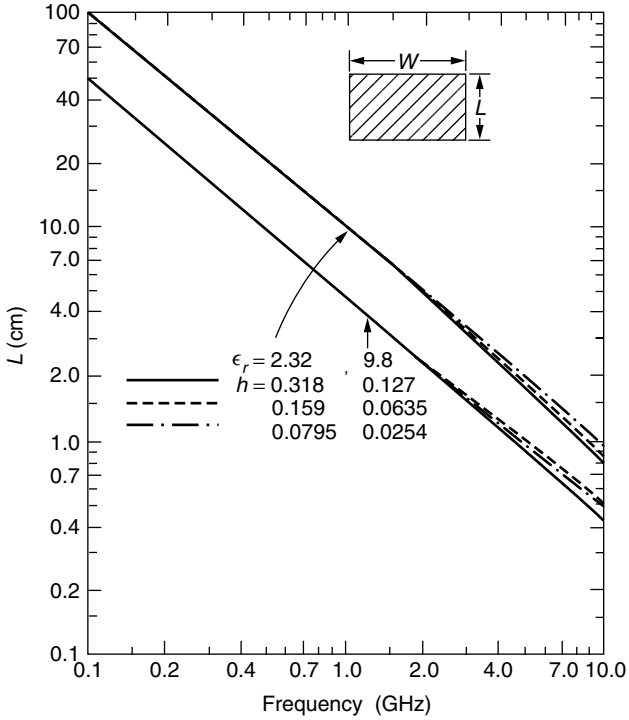


Figure 6. Element length versus frequency for different dielectric substrates [8].

2.3.5. Input Impedance. The only free parameter left is the location of the feeding point. Using this parameter, one can design a patch with almost any input impedance at resonance. Because of its fundamental current distribution, a patch fed in the center has a zero-ohm input impedance. To avoid the excitation of cross-polarization currents (for linear polarization), the feeding point has to be positioned symmetrically with respect of the width of the patch. Figure 7 shows the dependence of the patch input impedance on the location of the feeding point.

2.3.6. Q Factor and Losses. The quality factor of the patch is given by

$$Q = \frac{Q_r R_T}{R_r} \tag{19}$$

where Q_r is the quality factor associated with the radiation resistance [10]:

$$Q_r = \frac{c\sqrt{\varepsilon_e}}{4f_r h} \tag{20}$$

$$R_c = 0.00027\sqrt{f_r}\frac{L}{W}Q_r^2 \tag{21}$$

$$R_d = \frac{30\tan\delta}{\varepsilon_r}\frac{h\lambda_0}{LW}Q_r^2 \tag{22}$$

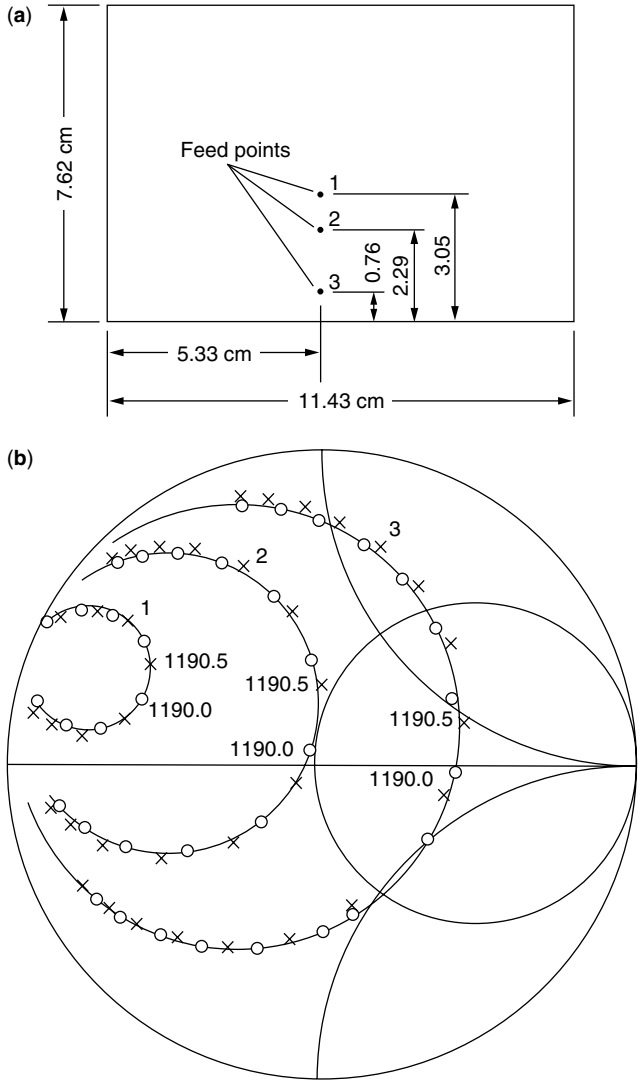


Figure 7. Experimental and theoretical loci for a microstrip rectangular patch antenna (xxx theoretical, ooo experimental) [9].

and

$$R_T = R_r + R_d + R_c \tag{23}$$

The radiation efficiency is thus

$$\eta\% = 100\frac{R_r}{R_T} \tag{24}$$

Figure 8 shows the radiation resistance R_T as a function of frequency for different dielectric substrates, and Fig. 9 shows the efficiency, $\eta(\%)$ as a function of frequency for different dielectric substrates.

2.3.7. Bandwidth. The bandwidth of the microstrip for VSWR < $\text{VSWR}_{\max}$ is given by

$$BW = \frac{\text{VSWR}_{\max} - 1}{Q_T\sqrt{\text{VSWR}_{\max}}} \tag{25}$$

This formula establishes two basic principles in the design of microstrip antennas; for a certain frequency,

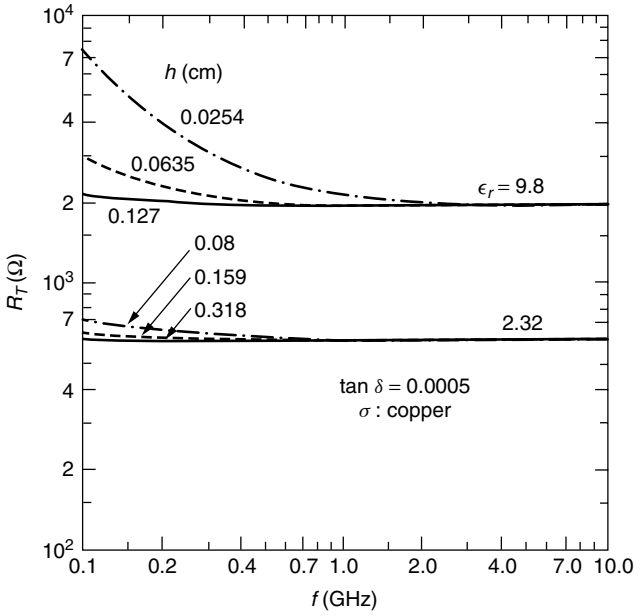


Figure 8. Radiation resistance as a function of frequency for different dielectric substrates [8].

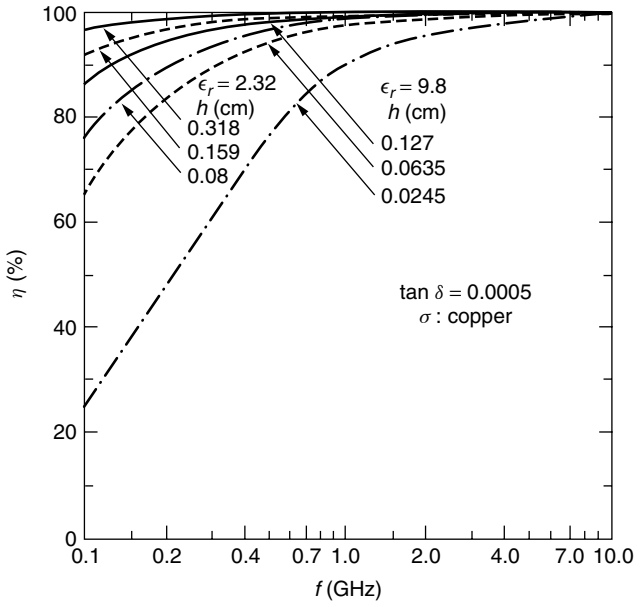


Figure 9. Efficiency as a function of frequency for different dielectric substrates [8].

the bandwidth of the antenna is directly proportional to the substrate thickness and inversely proportional to the dielectric constant of the substrate.

2.3.8. Directivity and Gain. Bahl and Bhartia [8] showed that a good approximation for the directivity is given by

$$D \cong 6.6 \quad W \ll \lambda_0 \tag{26}$$

$$D \cong \frac{8W}{\lambda_0} \quad W \gg \lambda_0 \tag{27}$$

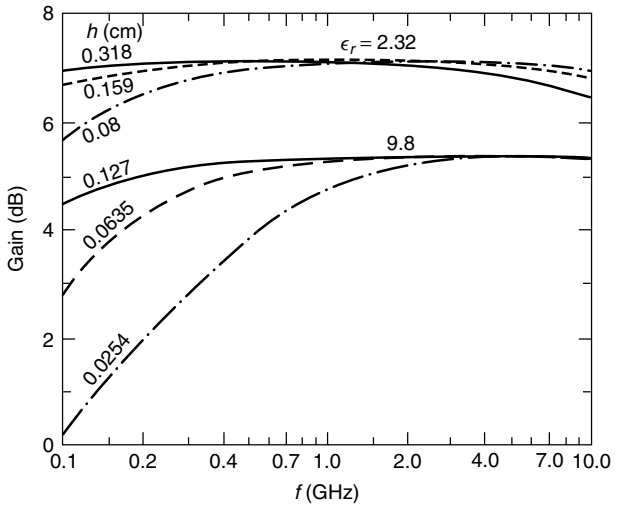


Figure 10. Gain as a function of frequency for various rectangular microstrip antennas [8].

For any antenna, the gain is defined as $G = \eta D$. Figure 10 shows the Gain as a function of frequency for various rectangular microstrip antennas.

2.4. The Impact of Manufacturing Tolerance on the Electrical Characteristics of the Rectangular Patch Antenna

A number of facts have to be considered in the fabrication of microstrip antennas:

1. Substrates with a low dielectric constant have variation in the dielectric constant of about $\pm 1\%$ and $\pm 5\%$ in thicknesses.
2. The tolerances for higher dielectric constant substrates are about $\pm 2\%$ (for dielectric constant) and $\pm 4\%$ for thickness.
3. A significant source for errors is in the etching process. A proper fabrication process has to include good control on the surface quality of the substrate, and adequate metalization thickness.

Bahl and Bhartia [8] present formulas for the sensitivity of some of the parameters discussed above. These formulas are reproduced below:

1. The change in the resonance frequency as a function of the variation of the length of the patch and the effective dielectric constant of the substrate

$$|\Delta f_r| = \sqrt{\left(\frac{\partial f_r}{\partial L}\Delta L\right)^2 + \left(\frac{\partial f_r}{\partial \varepsilon_e}\Delta \varepsilon_e\right)^2} \tag{28}$$

2. The change in effective dielectric constant of the substrate as a function of the variation of the width of the patch (W), the thickness of the substrate (h), the dielectric constant of the substrate (ε), and the thickness of the metalization (t) is

$$|\Delta \varepsilon_e| = \sqrt{\begin{array}{l}\left(\dfrac{\partial \varepsilon_e}{\partial W}\Delta W\right)^2 + \left(\dfrac{\partial \varepsilon_e}{\partial h}\Delta h\right)^2 \\ + \left(\dfrac{\partial \varepsilon_e}{\partial \varepsilon_r}\Delta \varepsilon_r\right)^2 + \left(\dfrac{\partial \varepsilon_e}{\partial t}\Delta t\right)^2\end{array}} \tag{29}$$

From the previous three equations we can derive the formula for the relative change in the resonance frequency:

$$\frac{|\Delta f_r|}{f_r} = \sqrt{\begin{array}{l}\left(\dfrac{\Delta L}{L}\right)^2 + \left(\dfrac{0.5}{\varepsilon_e}\right)^2 \left\{\left(\dfrac{\partial \varepsilon_e}{\partial W}\Delta W\right)^2 + \left(\dfrac{\partial \varepsilon_e}{\partial h}\Delta h\right)^2 \right. \\ \left. + \left(\dfrac{\partial \varepsilon_e}{\partial \varepsilon_r}\Delta \varepsilon_r\right)^2 + \left(\dfrac{\partial \varepsilon_e}{\partial t}\Delta t\right)^2\right\}\end{array}} \tag{30}$$

Figure 11 shows the variation of change in the fractional resonant frequency of a rectangular microstrip antenna with frequency for $\varepsilon_r = 2.32$ and given tolerances. The complete design of a microstrip antenna has to factor also in polarization, frequency response (wideband, multiband), and feeding mechanism. The following sections address these issues in a broader context and present a large variety of geometries.

3. FEEDING METHODS FOR MICROSTRIP ANTENNAS

3.1. Microstrip Antenna Configurations

As they are essentially printed circuits, microstrip antennas can have different geometric shapes and dimensions. Since the early 1970s, numerous configurations have been proposed [11]. Some of these geometries are shown in Fig. 12. The electrical characteristics of these shapes are somewhat similar, all having a broadside beam generated by a fundamental mode. The slight difference in the physical area occupancy or multimode (or higher-mode) operation might make one geometry more appropriate for certain applications. All these geometries, however, can each be fed in similar ways.

3.2. Coaxial Feed

The coaxial feeding method is mostly appropriate for single elements. The location of the feeding point determines

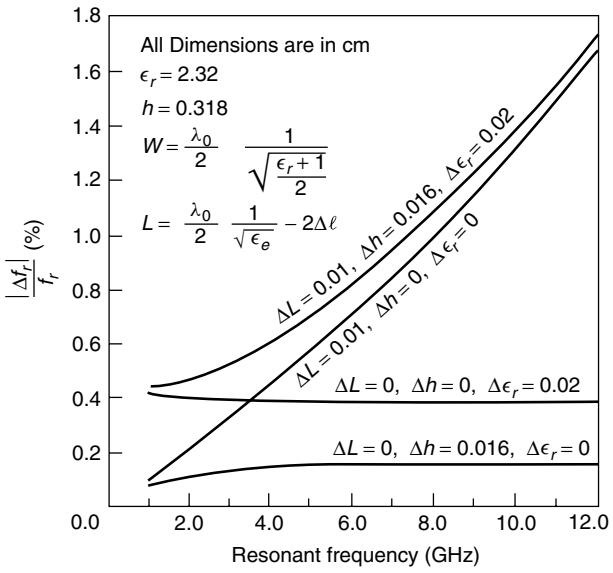


Figure 11. Variation of change on the fractional resonant frequency of a rectangular microstrip antenna with frequency for $\varepsilon_r = 2.32$ and given tolerances [8].

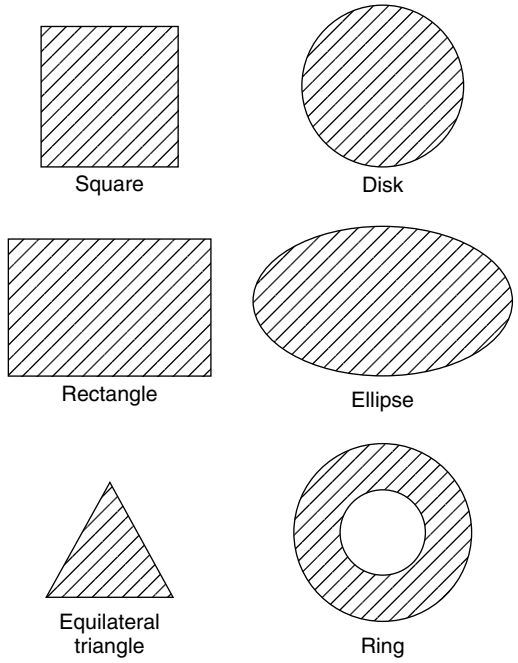


Figure 12. Basic microstrip patch antenna shapes commonly used in practice [8].

the input impedance and the polarization. The input impedance calculations of the single element will have to include besides the patch self-impedance a serial reactance component as shown in Eq. (12) The geometry of the coaxial fed microstrip patch is shown in Fig. 13. The central pin of the coaxial feed is connected to the patch at the "feeding point," and the shield of the coaxial feed is connected to the ground plane. A number of various coaxial-fed patch geometries are shown in Fig. 14.

3.3. Microstrip Line Feed

The coaxial-fed patch is not easy to array. In arrays, the most common way to feed the radiating element is using a microstrip line, which generally is an extension of the feed network.

The most used types of microstrip feeds are

1. The coplanar microstrip feed (Fig. 15)
2. The proximity (electromagnetic) coupled microstrip feed (Fig. 16)
3. The aperture coupled microstrip feed (Fig. 17)

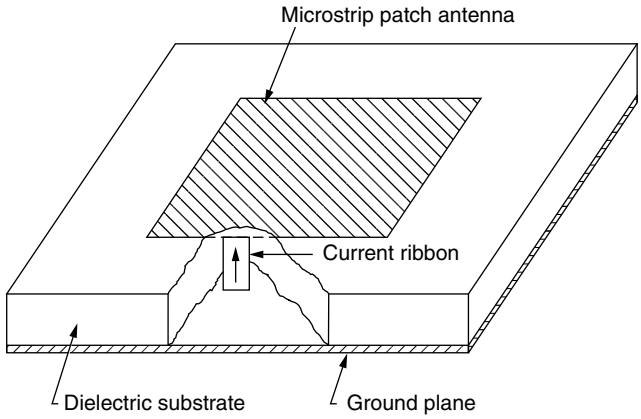


Figure 13. The coaxial-fed microstrip patch [8].

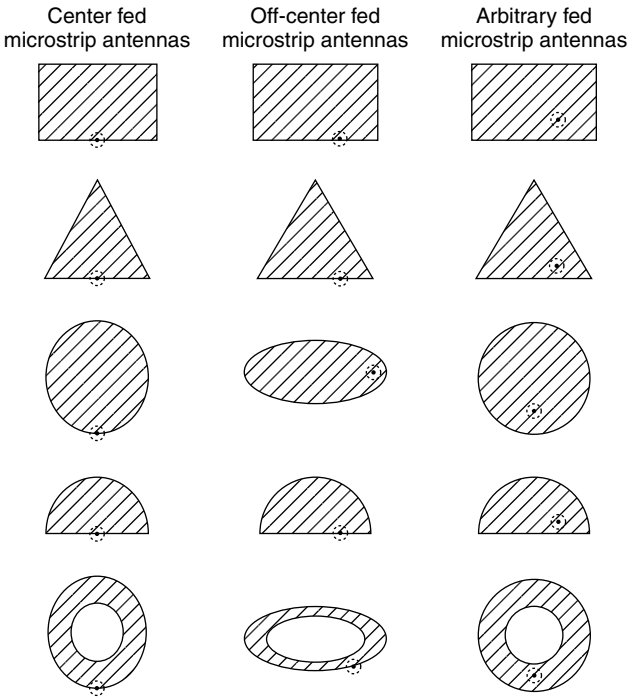


Figure 14. Coaxial fed microstrip antennas [8].

Figure 15 shows a number of variations of the coplanar microstrip feed: edge feed (a), gap feed (b), and inset feed (c).

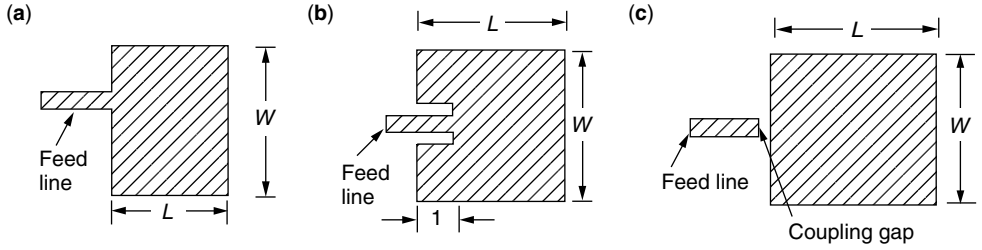


Figure 15. The coplanar microstrip feed.

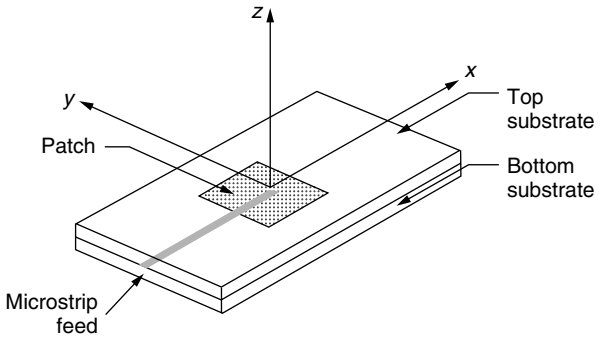


Figure 16. The proximity coupled microstrip patch.

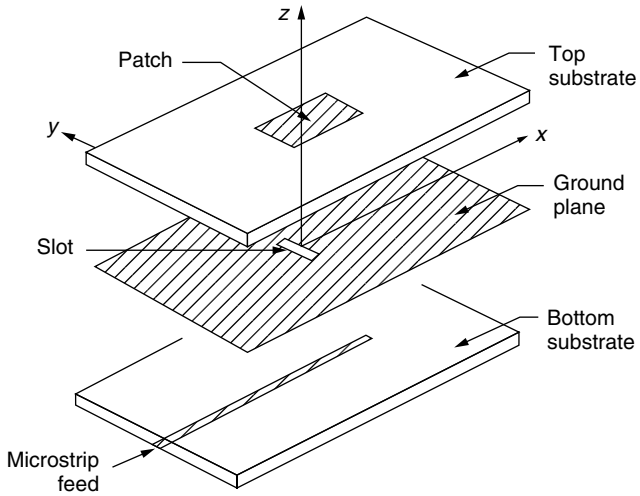


Figure 17. The aperture-coupled microstrip feed.

The patch impedance required to match the feed network determines the choice of any of the three. The patch self-impedance at the edge is typically high, and if 50 Ω is required, then the inset feed will have to reach inside the patch to a point where the patch self-impedance is 50 Ω. The coplanar microstrip feed has the advantage that it is printed on the same substrate as the radiating elements. In some cases, the design of such arrays might be difficult since the patches themselves might occupy most of the space on the substrate.

The mutual coupling between the radiating elements and the feeding network as well as the feeding network itself can create spurious radiation that might affect the overall performance of the array.

In order to avoid a crowded design, two substrates can be used, one for the radiating elements and one for the feeding network. In this case, the patches can be fed by a microstrip line sharing the same ground plane as the patch but located in between the patch and the ground plane (proximity feed). This feeding mechanism solves the "real estate" problem on the substrate; however, it does not address the spurious radiation problem. The total separation between the radiation of the radiating elements and the spurious radiation of the feeding network is achieved by using the aperture-coupled feeding method.

A comprehensive overview of the different type of feeding mechanisms is given by James and Hall [11] and Garg et al. [12].

4. POLARIZATION PROPERTIES OF MICROSTRIP ANTENNAS

4.1. Linear Polarization

In general, any rectangular or circular microstrip antenna fed in a symmetric way with respect to one axis will be linearly polarized. The difference between the different methods of feeding mentioned in the previous paragraphs is in the cross-polarization level. The probe feeding of a patch is symmetric in the H plane; however, it is not symmetric in the E plane, and this results in the excitation of cross-polarization currents. In addition, a thicker substrate implies a longer probe that radiates, and increases even more the cross-polarization level. The aperture-fed patch is symmetrically fed in both planes, and no cross-polarization currents are excited.

4.2. Circular Polarization

4.2.1. Singly Fed Circularly Polarized Microstrip Antennas. Traditionally, the singly fed circularly polarized microstrip antennas are very narrowband, both in terms of VSWR and axial ratio. A number of different geometries are shown in Fig. 18.

In general, circular polarization is obtained by superimposing two orthogonal current modes that are excited with equal amplitude and a phase differential of 90°. This can be achieved by introducing a perturbation segment, which excites a specific current distribution, consisting of

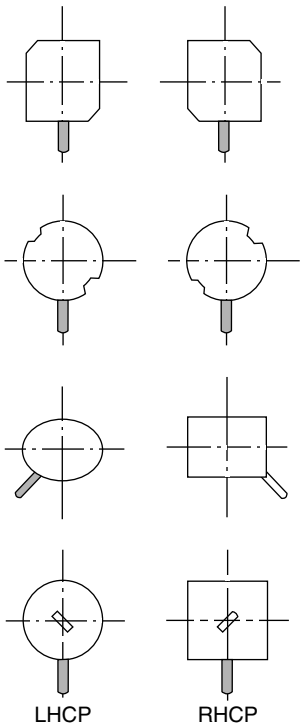


Figure 18. Singly fed circularly polarized patches [11].

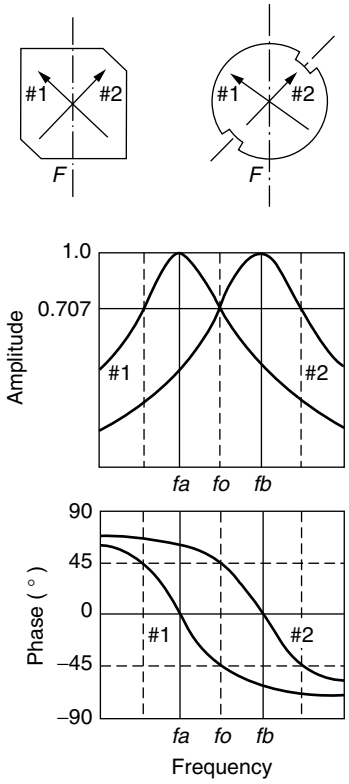


Figure 19. Amplitude and phase diagrams for singly fed circularly polarized microstrip antennas [11].

the two modes, resonant at slightly different frequencies. At the central frequency, the modes self-impedance fulfills the condition mentioned above (Fig. 19). This effect though, for single layered patches is very narrowband [6]. The advantage of the singly fed CP microstrip antennas is that they are easy to array, and in terms of array topology, they are similar to the linearly polarized version. Owing to their narrowband characteristics, they have few applications. For stacked patched however, a wider band can be achieved [13]. A detailed design procedure for the singly fed circularly polarized microstrip antennas is given in Chapter 4 of Ref. 11.

4.2.2. Dual-Fed Circularly Polarized Microstrip Antennas. When a wider band of operation is required (for VSWR and axial ratio), the dual-fed configuration is a better choice (Fig. 20). In this case, the excitation of the appropriate modes is done outside the radiating element. As shown in Fig. 20, the overall size of the element (which now includes the circuit generating the circular polarization) is considerably larger. In an array, a large element would force a large separation between elements, resulting in grating lobes.

An ingenious solution for the tradeoff between bandwidth and element size in arrays was presented by John Huang [14] (Fig. 21). The idea consists of creating circularly polarized subarrays from linearly polarized elements. This sequential feeding scheme allows for an excellent circular polarization over a relatively wide frequency bandwidth. Moreover, the array is capable of scanning in the principal planes to relatively wide angles from

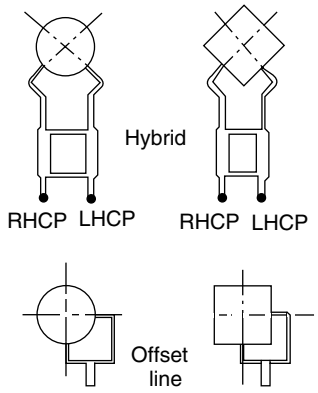


Figure 20. Dual-fed CP patches [11].

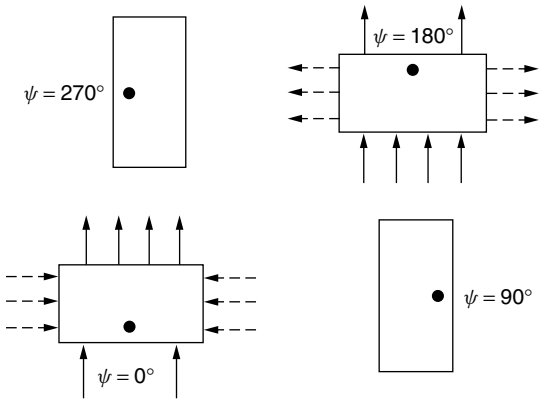


Figure 21. A 2×2 microstrip subarray that generates CP with LP elements [14].

its broadside direction without significant degradation to the axial ratio. This idea was developed further using singly fed circularly polarized instead of linearly polarized elements [15,16].

5. BANDWIDTH CHARACTERISTICS OF MICROSTRIP ANTENNAS

5.1. Introduction

The basic, single-layer microstrip antennas are $2\frac{1}{2}$D structures, and therefore, electrically very small. Since the early 1970s, many variations of the elementary patch were developed: multilayer microstrip antennas (or stacked patches), tall patches, cluster patches, and slotted patches. In essence, all these variations have as a goal the realization of radiating elements, which can be easily arrayed, with a radiation pattern similar to the elementary patch and that, finally, have a significantly larger bandwidth than the original microstrip patch antenna. All the techniques mentioned above effectively increase the electrical volume of the radiating elements, and generate radiating elements with a lower Q.

The difference between these techniques resides in the different tradeoffs they present, such as bandwidth versus physical volume, manufacturing cost, cross-polarization, and radiation pattern shape.

5.2. The Single Patch

The choice of the different parameters in the design of the single patch offers *some* latitude, even though quite limited, in the bandwidth characteristics:

1. The thicker the substrate, the wider the bandwidth.
2. The lower the dielectric constant of the substrate, the wider the bandwidth.

When the dielectric substrate is too thick, the efficiency and the crosspolarization of the antenna can be of concern; the surface wave dependency on the substrate thickness *is not* monotonic, and an excellent study of this phenomenon is given by Pozar [17]. However, when the substrate is excessively thick (allowing for the excitation of higher-order surface waves), the efficiency is considerably affected. Figures 22 and 23 show the surface wave efficiency as a function of the substrate thickness for two dielectrics, $\varepsilon_r = 2.55$ and $\varepsilon_r = 12.8$, respectively.

The losses due to the dielectric heating *are* monotonic and inversely proportional to the substrate thickness (Fig. 24). In addition, for a single-layer patch *resonating at a certain frequency*, by increasing the dielectric substrate, the directivity of the antenna is reduced. This is due to the fact that the antenna is smaller.

A special class of tall patches is the suspended patches. Rather than printing the patch on a grounded substrate,

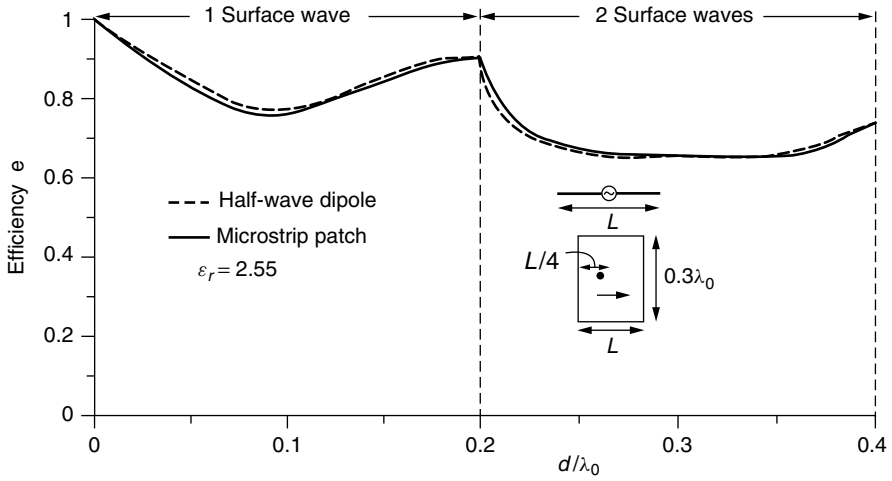


Figure 22. Loss due to surface wave for a half-wave printed dipole and a microstrip patch versus the substrate thickness for $\varepsilon_r = 2.55$, with patch width $0.3\lambda_0$ [17].

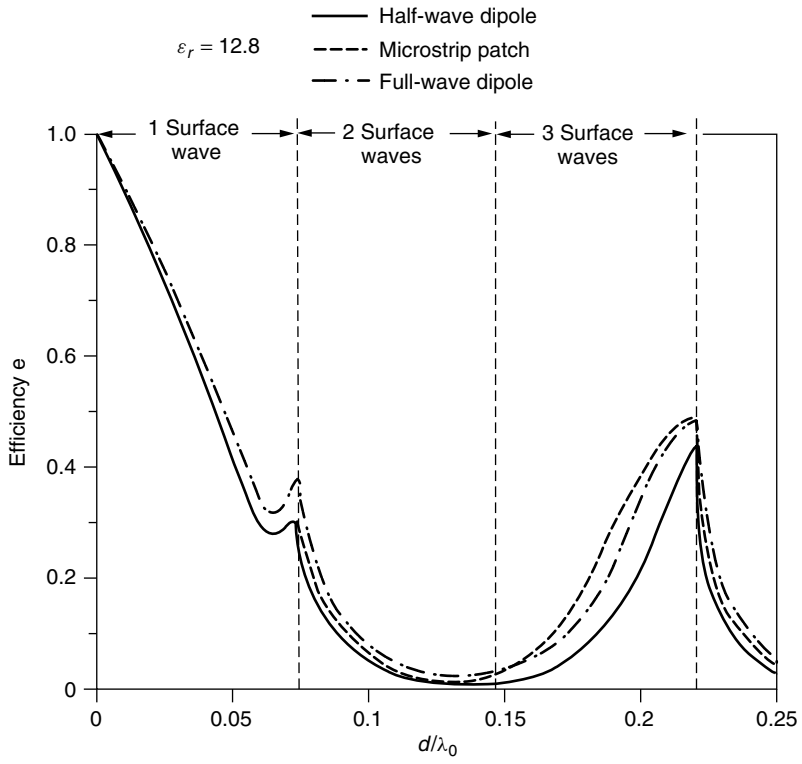


Figure 23. Loss due to surface wave for a half-wave printed dipole and a microstrip patch versus the substrate thickness for $\varepsilon_r = 12.8$, with patch width $0.15\lambda_0$ [17].

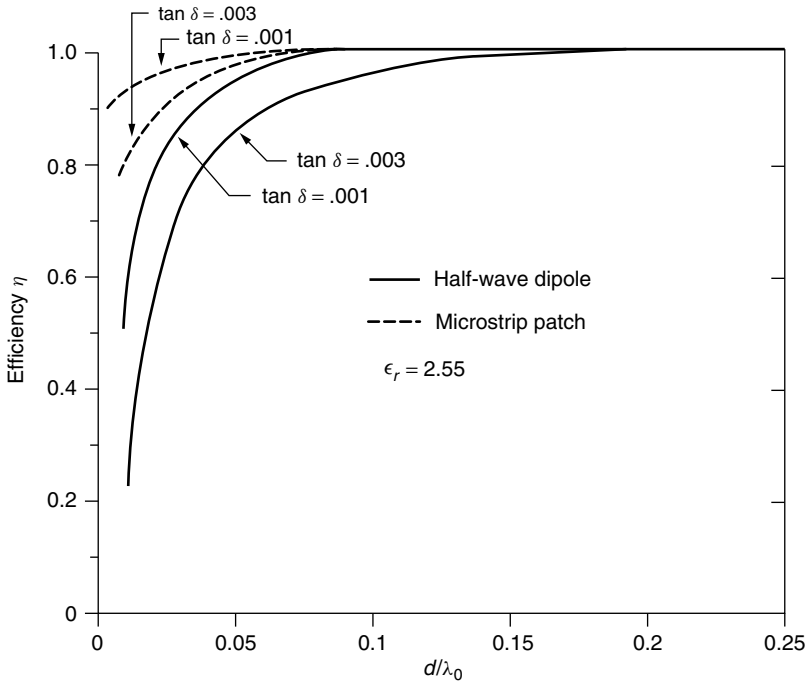


Figure 24. Loss due to dielectric for a half-wave printed dipole and a microstrip patch versus the substrate thickness for $\varepsilon_r = 2.55$, with patch width $0.3\lambda_0$ [17].

the patch is made of bare metal (or printed on a very thin dielectric substrate, typically 2–5 mils) and separated from the ground either by foam or by a supporting standoff. The widest bandwidth (in terms of VSWR) for a suspended patch reported in the literature is 95% [18].

The design reported in Ref.18 incorporates three principles of bandwidth enhancement.

1. Large separation between the patch and the ground plane
2. Low dielectric (air)
3. Multiresonant patch geometry

The geometry is shown in Fig. 25, and theoretical prediction and measurements are compared in Fig. 26.

The shape of the patch is such that different parts of the patch are resonant at different frequencies [19]. The distance between the radiating element and the ground plane is about $\lambda/4$ at midband, and using a $\lambda/4$ probe to feed the patch would allow the probe to radiate like a monopole. This is why a 3D transition was used to feed the patch. At least for 45% of the band, the cross-polarization of the element is better than 10 dB within the −3-dB beamwidth. On broadside, the cross-polarization is better than 30 dB. As shown in Fig. 26, the VSWR is less than 2 from 2.2 to 4.3 GHz. At frequencies higher than about 3.5 GHz, the 3D transition itself is radiating, and generates a high level of cross-polarization.

This example emphasizes the fact that the term *bandwidth* has to be carefully defined when referring to antenna performance; sometimes the *VSWR bandwidth* is different from the *cross-polarization bandwidth, directivity bandwidth, axial ratio*, and other parameters.

5.3. Nonresonant Methods for Bandwidth Enhancements

The simplest way to improve the frequency response of a microstrip antenna is to use a matching network. In Refs. 20 and 21, 10–12% bandwidths are reported for a relatively thin patch, using a lossless matching network. Lossless matching networks, though, have only a limited impact, and in some cases they might occupy too much

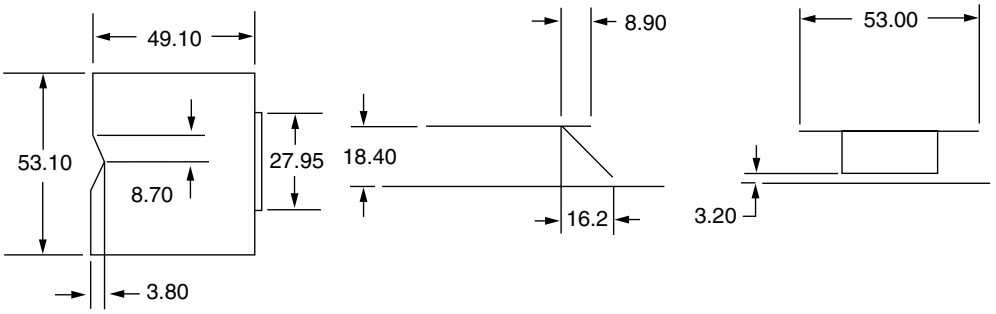


Figure 25. Dimensions (in millimeters) for the wideband microstrip single-layer patch antenna [18].

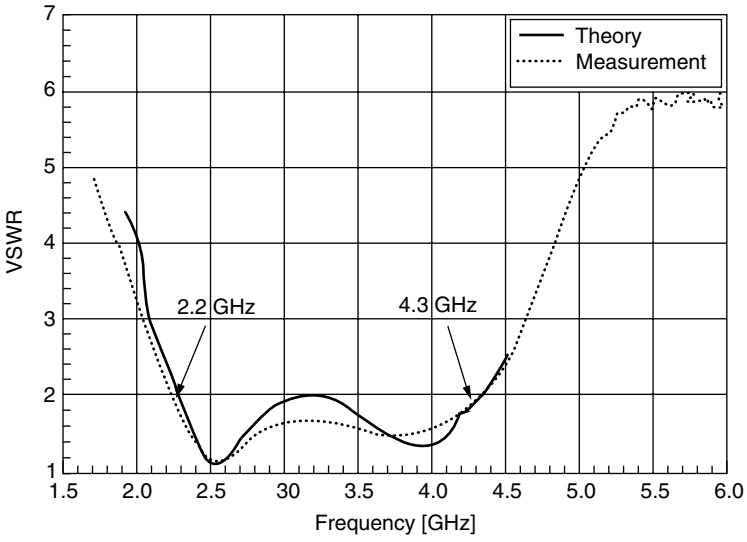


Figure 26. The VSWR of a wideband microstrip single-layer patch antenna [18].

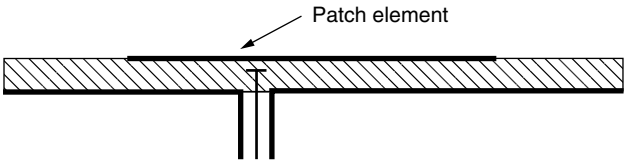


Figure 27. Capacitive feeding of a single patch.

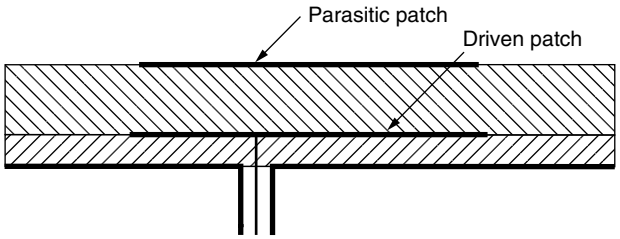


Figure 28. Bandwidth improvement using stacked patches.

space on the board. In addition, if the matching network were too complex, they would create spurious radiation.

A simple matching technique is capacitive feeding (Fig. 27). The matching is achieved by controlling the size of the tab and its distance from the patch.

5.4. Multiresonator Microstrip Antennas

5.4.1. The Stacked Patch. The single patch can be considered as a resonator. By adding an additional patch (Fig. 28), an additional resonator is created. By setting the resonance dimensions of the driven patch and the parasitic patch appropriately, the broadband or dual-band effect can obtained. The physical interpretation (or the equivalent circuit) of such a structure is extremely difficult to generate, mainly because of the mutual coupling between these two resonators. Therefore, only full-wave modeling can provide a good prediction of the electrical characteristics of this antenna [22,23]. Table 1 gives an idea of the bandwidths that can be achieved [24].

The case of the stacked patches is special in the sense that the radiating element has almost the same size (in the substrate plane) as the single patch itself, and it does not require additional space. The two patches have to be very close in size to obtain the broadband effect. The other methods for bandwidth enhancement, described below, involve larger elements and/or some price to pay in performance (front-to-back design, cross-polarization, complexity of fabrication, etc.).

Intuitively, the next step would be to add more parasitic elements. Since the excitation of the parasitic element is by coupling, the broadband effect is lost very quickly.

5.4.2. Coplanar Parasitic Elements. A different way to use parasitic elements is in the coplanar configuration. Figure 29 shows the geometry of a probe-fed patch

Table 1. Experimental Results for Stacked Two Layer Antennas [21]

Antenna Geometry	Frequency Band	Bandwidth (%)	Beamwidth H-Plane (Degrees)	Gain (dbi)	Sidelobe Levels H-Plane (dB)	Polarization
Circular disk	S	15	72	7.9	−22	Linear
Circular annular disk	S	11.5	78	6.6	−14	Linear
Rectangular	S	9	70	7.4	−25	Linear
Square	S	9	72	7	−22	Linear
Circular disk	S	10	72	7.5	−22	Circular
Circular disk	X	15	72	7.5	−25	Circular

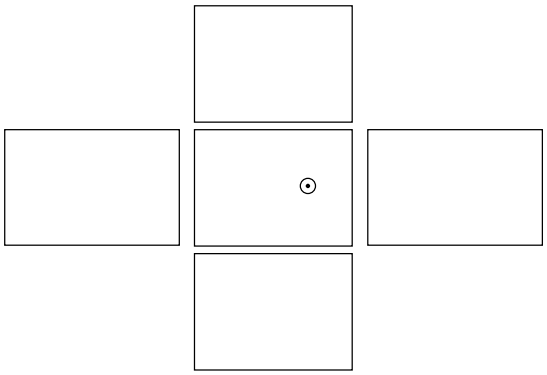


Figure 29. A probe-fed patch, with four edge-coupled parasitic elements [25].

feeding four coplanar parasitic patches. In order to obtain the appropriate level of coupling, the gap between the patches has to be very small. That requires very tight tolerances in fabrication, which might be difficult to achieve. Bandwidths up to 25% have been reported [25–26] (Fig. 30); however, control over the shape of the beams might be difficult (Fig. 31). Because of the tight tolerances required in the fabrication of edge-coupled elements, direct coupling was proposed [27]. Figure 32 shows the proposed geometry. As shown in Fig. 33, the experimental bandwidth is 810 MHz (24% at $f_o = 3.38$ GHz), which is about 7.4 times the bandwidth of the typical rectangular patch antenna printed on the same substrate. The radiation patterns (Fig. 34), however, vary quite significantly across the operating frequency band, which might be unacceptable in some applications.

5.4.3. Aperture-Coupled Microstrip Antennas. The basic configuration of the aperture-coupled microstrip antenna is shown in Fig. 17. Initially developed as a way

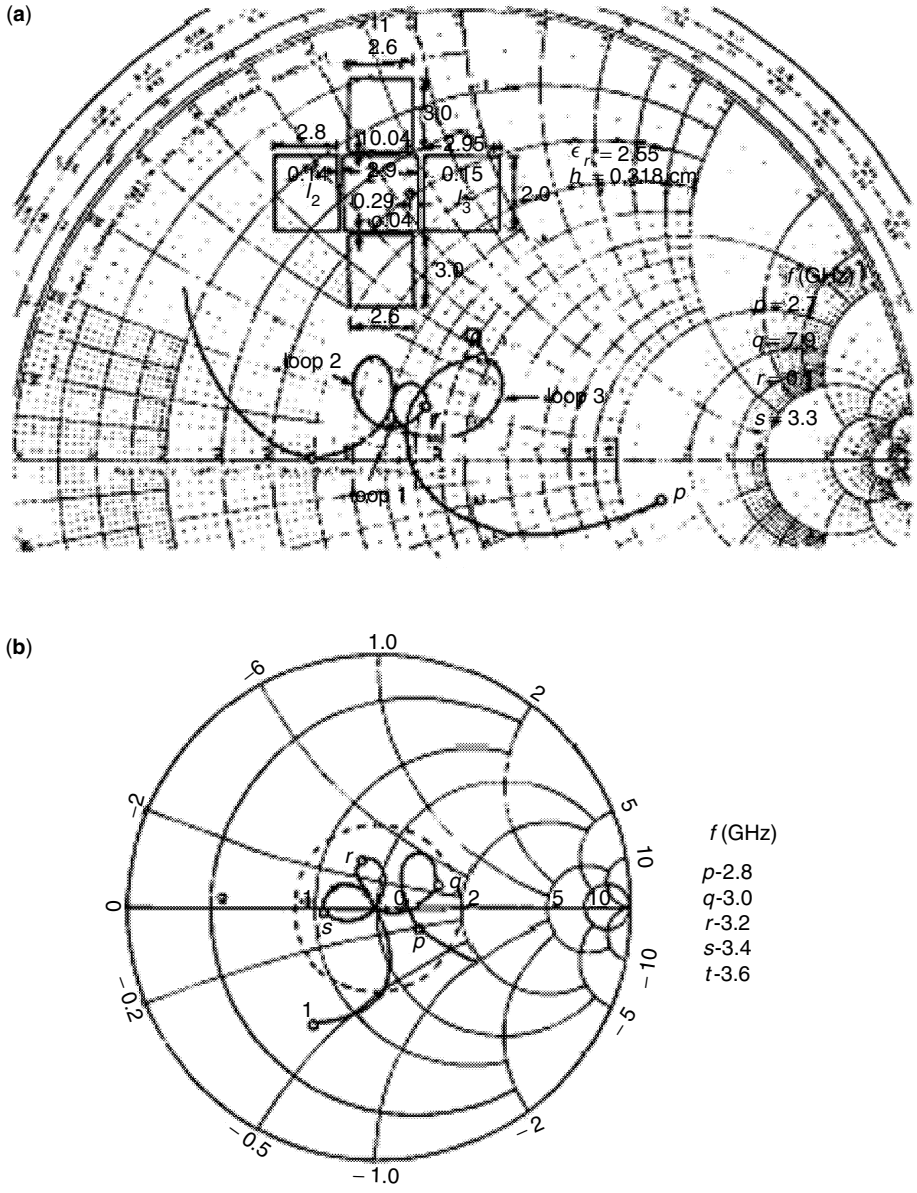


Figure 30. (**a**) Theoretical input impedance locus of FEGCOMA shown in inset and (**b**) experimental input impedance locus of FEGCOMA with modified dimensions [25].

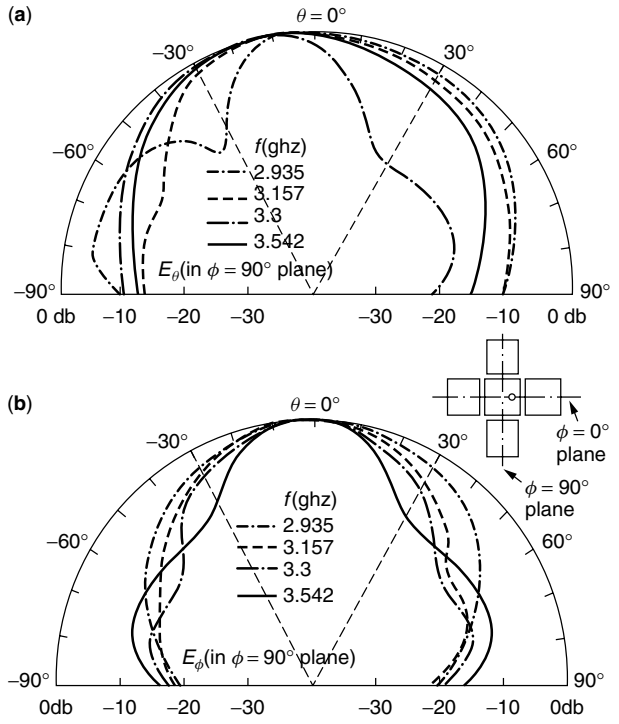


Figure 31. Experimental values of the radiation fields of FEGCOMA [25]: (**a**) E_θ in $\phi = 0^\circ$ plane and (**b**) E_ϕ in $\phi = 90^\circ$ plane.

to separate the feeding line from the radiating element, the aperture-coupled patch introduces an additional resonator: the coupling aperture. To avoid back radiation, the coupling aperture should not be resonant; however, its resonance can be *close* to the patch resonance, so that the antenna bandwidth is slightly increased. A number of different geometries based on the aperture-coupled microstrip antenna were developed:

1. The aperture-coupled coplanar dipole array [28] is shown in Figs. 35 and 36. Croq and Pozar [28] discuss only the multiband case; however, this geometry is conceivably appropriate for broadband applications.
2. The aperture-coupled stacked patch antenna, which, as reported [29,30], can achieve 50% bandwidth, is shown in Fig. 37. When using the aperture to feed the radiating elements, usually the tradeoff is between bandwidth and the amount of back radiation allowed. Some attempts were made to suppress the backradiation; a shielding plane was placed behind the antenna. While the back radiation is reduced, the shielding plane allows for the excitation of parallel-plate modes, which can seriously degrade the efficiency of the antenna. Furthermore, this bandwidth enhancement is done at the expense of much greater manufacturing complexity.

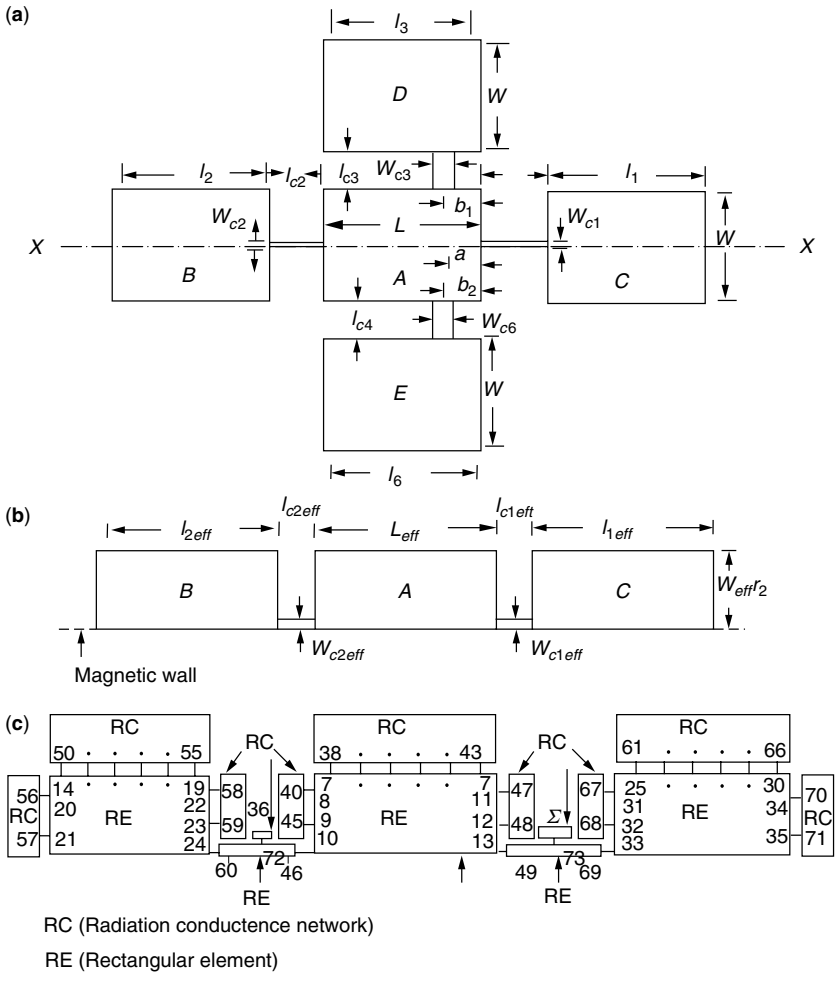


Figure 32. (**a**) Four edges directly coupled microstrip antenna (FEDCOMA) [26]; (**b**) even-mode half-section of REDCOMA; and (**c**) its segmented network.

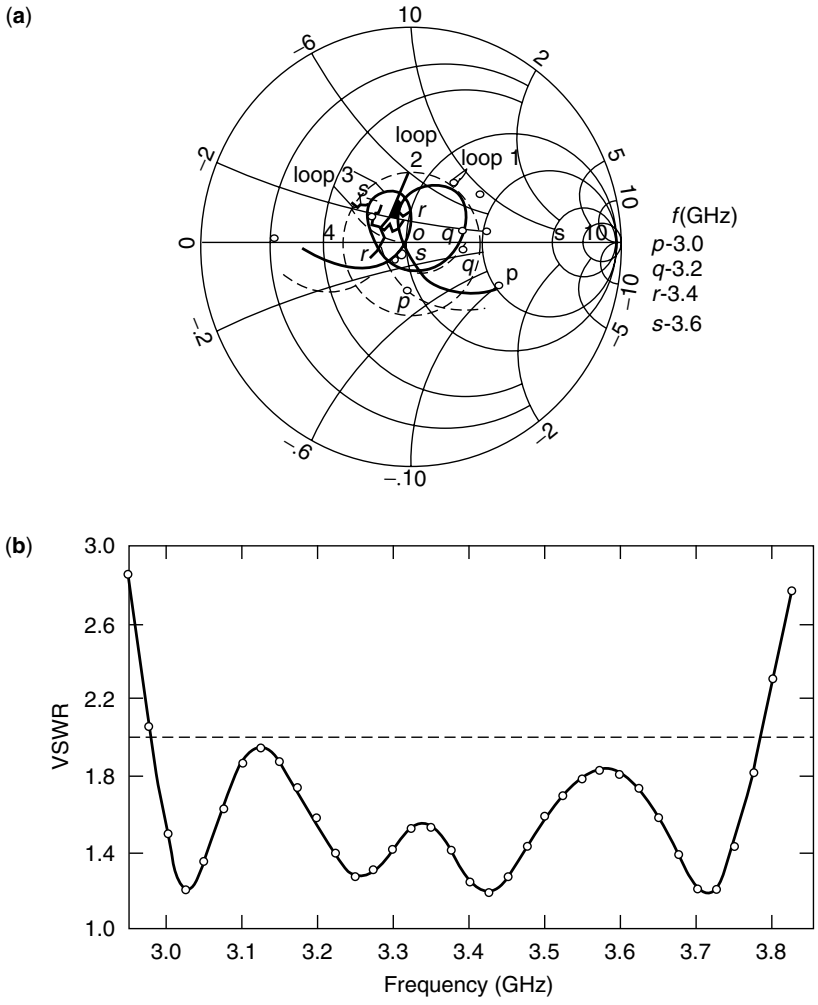


Figure 33. (**a**) Theoretical (---) and experimental (-o-) input impedance loci and (**b**) experimental VSWR variation with frequency of FEDCOMA [27].

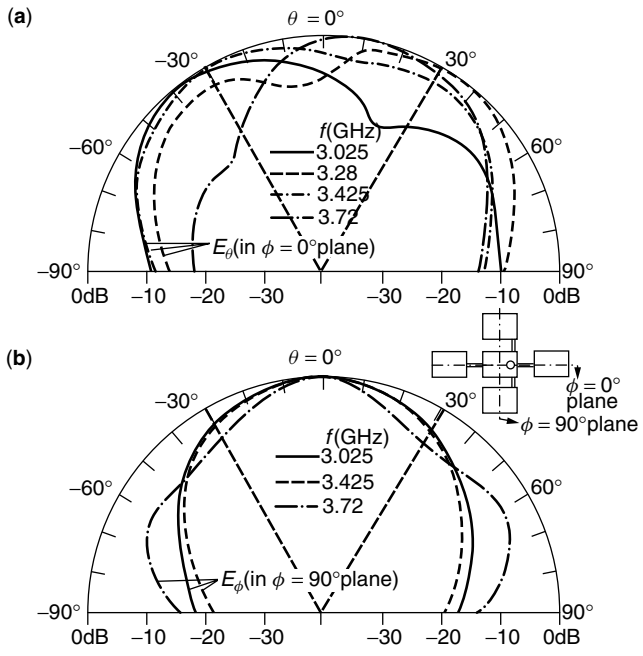


Figure 34. Experimental values of (**a**) E_θ in $\phi = 0^\circ$ plane and (**b**) E_ϕ in $\phi = 90^\circ$ plane of FEDCOMA [27].

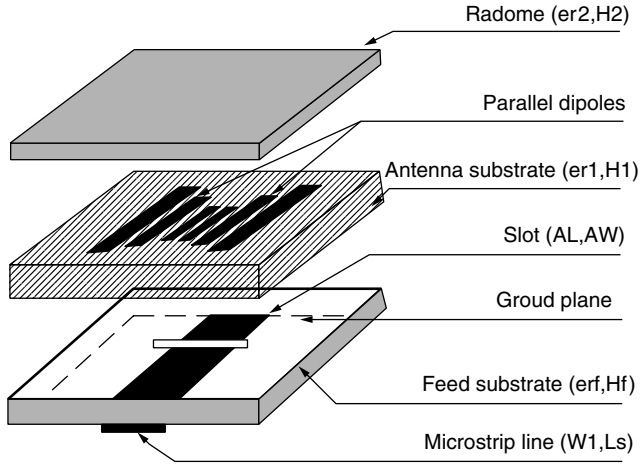


Figure 35. Multifrequency microstrip antenna composed of parallel dipoles aperture-coupled to a microstrip line [28].

6. MUTUAL COUPLING

In an array, the mutual coupling between elements can have a significant impact on the array radiation pattern as well as the input impedance. It can be either calculated

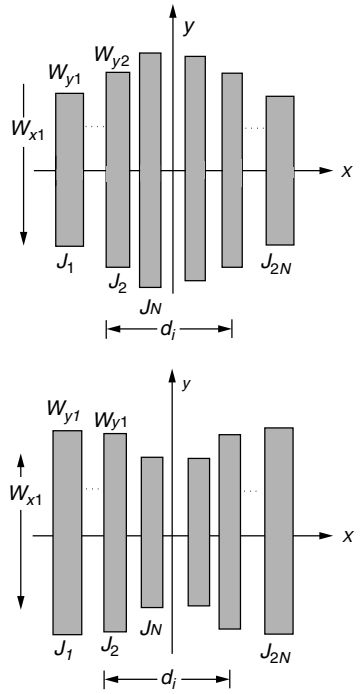


Figure 36. Two configurations of the multiple-resonator aperture-coupled antenna: (**a**) MFR1 and (**b**) MFRZ [28].

or measured, so that the feed network design can be modified to compensate (where possible) for its effect. In very large scanning arrays, the mutual coupling can create *blindness*, a situation in which the input reflection coefficient is very close to 1. The blindness effect will be discussed in Section 7.

The mutual coupling mechanism in microstrip antennas consists of two components: the radiation and the surface waves. A full-wave analysis of the mutual coupling between rectangular microstrip antennas is presented by Pozar [32] (Fig. 38). When the patches are very close (less than about $\lambda/10$), the *H*-plane coupling is slightly stronger than the *E*-plane coupling; however for larger distances, the *E*-plane coupling is significantly stronger, due to the excitation of surface waves. The mutual coupling depends on the substrate and the shape of the patch. A measurement study is presented by Jedlicka et al. [33]. Figure 39 shows the effect of the mutual coupling on its input impedance, and Figs. 40–42 show the mutual coupling for the principal planes and different geometries.

7. MICROSTRIP ARRAYS

7.1. Introduction

The electrical characteristics of microstrip antenna elements were described above. However, they are even

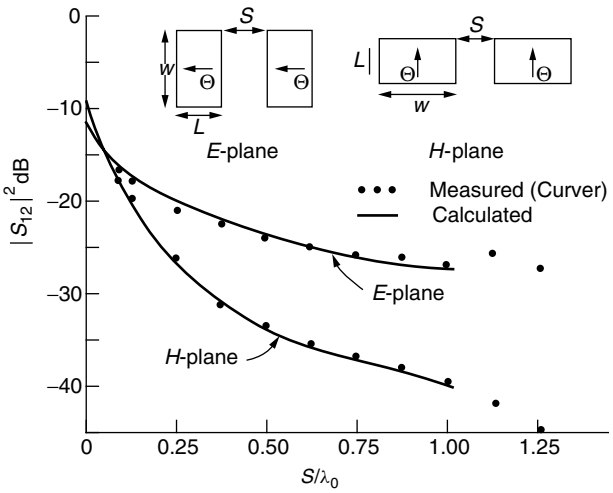


Figure 38. Measured and calculated mutual coupling between two coaxial-fed microstrip antennas for both *E*-plane and *H*-plane coupling ($W = 10.57$ cm, $L = 6.55$ cm, $d = 0.1588$ cm, $\varepsilon_r = 2.55$, $f_0 = 1410$ MHz) [32].

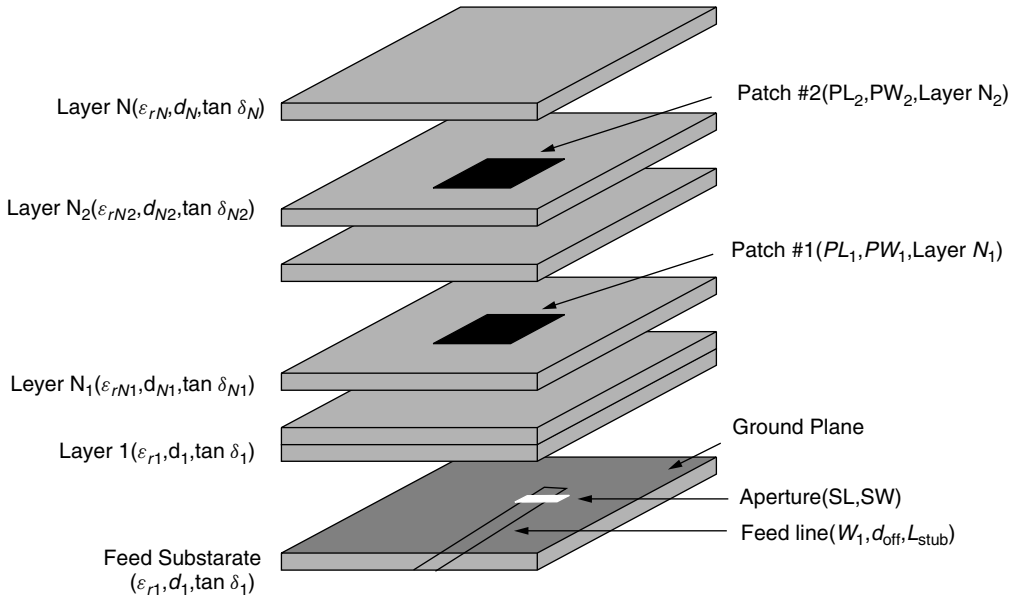


Figure 37. The wideband aperture-coupled stacked patch microstrip antenna [30].

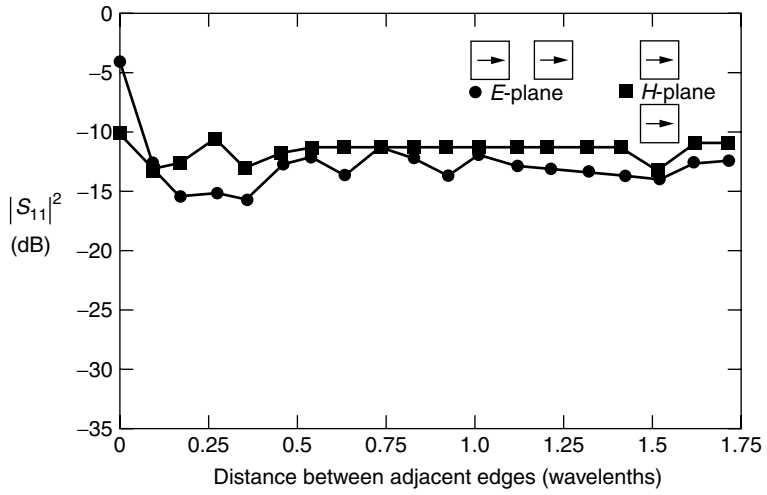


Figure 39. Measured $|S_{11}|^2$ values at 1410 MHz for 10.57 cm (radiating edge) × 6.55 cm rectangular patches with 0.1575 cm substrate thickness [33].

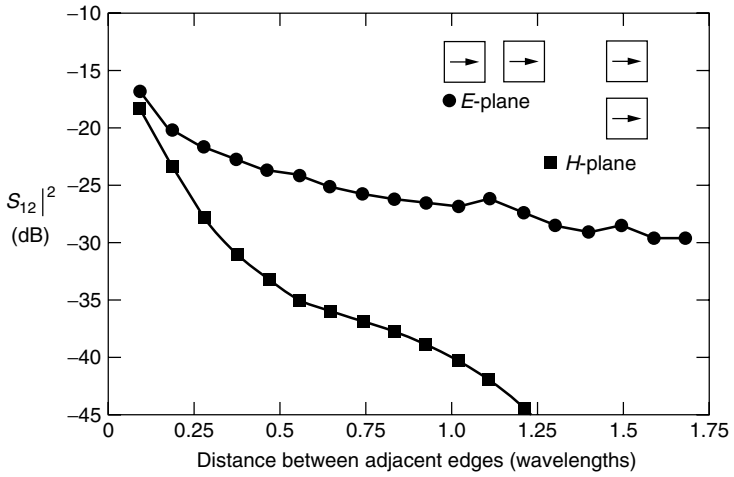


Figure 40. Measured $|S_{12}|^2$ values for the rectangular patch of Fig. 39 [38].

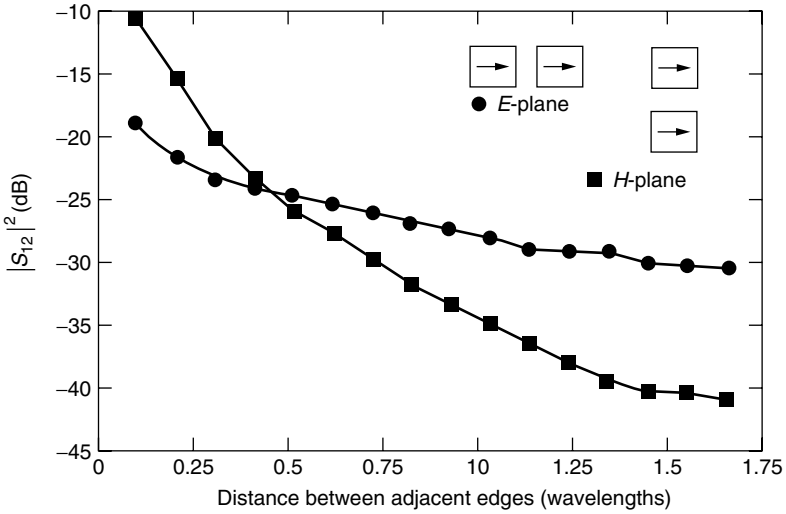


Figure 41. Measured $|S_{12}|^2$ values at 1560 MHz for 5.0 cm (radiating edge) × 6.0-cm nearly square patches with 0.305 cm substrate thickness [33].

more attractive in the array context. Compared to other types of arrays, microstrip arrays are relatively easy to manufacture, and are light and conformal. Since they essentially are printed circuits, they allow a significant freedom of design that results in a large variety of configurations: serial-fed arrays, parallel-fed arrays, and a significant number of different combinations between serial and parallel feeding techniques.

Array antennas in general and microstrip arrays in particular can be designed to have a fixed beam of a certain

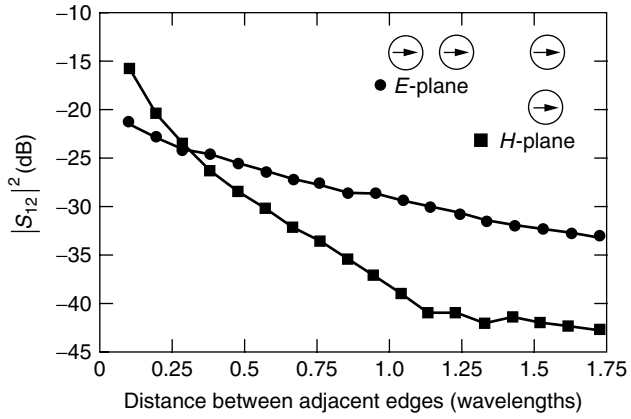


Figure 42. Measured $|S_{12}|^2$ values at 1440 MHz for circular patches with a 3.85-cm radius and a feed point location at 1.1 cm radius. The substrate thickness is 0.1575 cm [33].

shape or a beam that scans (using phase shifters or time-delay devices) or multiple beams (where the elements are fed by a special space feed).

7.2. Linear Arrays

By definition, linear arrays consist of radiating elements positioned at finite distances from each other along a straight line. In terms of the feed mechanism, linear arrays can be

1. Parallel-fed by a printed power divider
2. Serially fed with two-port radiating elements
3. Serially fed with one-port radiating elements

The parallel-fed array simply uses a power divider (usually printed on the same substrate as the radiating elements) to feed each radiating element (Fig. 43a). The junctions can be symmetric (for uniform amplitude) or asymmetric (e.g., for low-sidelobe design). In the case of low-sidelobe arrays, the bandwidth is determined not only by the bandwidth of the radiating elements but also by the bandwidth of the feed network. An important factor is played by the power dividers; the commonly used T junction does not have a very good isolation, and if the whole circuit is not very well matched, the amplitude and phase distribution generated will have errors. To alleviate this problem, Wilkinson power dividers could be used instead. When the electrical distances between the input port and all the other ports are identical, the phase distribution obtained is uniform and the beam generated is "squintless." The direction of the beam is independent of frequency and also of the spacing between elements.

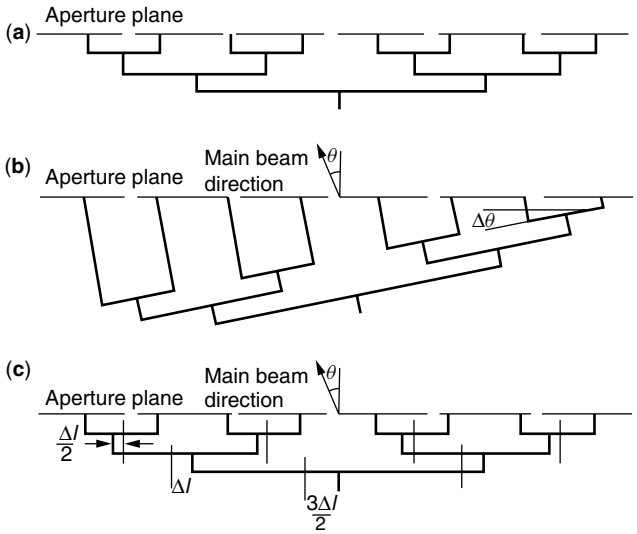


Figure 43. Parallel networks (**a**) without beam scan; (**b**) with beam scan, type 1; and (**c**) with beam scan, type 2.

A parallel feed network can be used to produce a scanning beam by simply using delay lines (Fig. 43b). The scanning angle is given by

$$\theta_0 = \arcsin\left(\frac{\delta}{2\pi}\frac{\lambda_0}{d}\right) \tag{31}$$

$$\delta = 2\pi\frac{\Delta l}{\lambda_t} \tag{32}$$

where δ = incremental phase difference between consecutive elements
d = distance between consecutive elements
λ_0 = wavelength in free space
λ_t = wavelength in transmission line
Δl = transmission line extension rate from one element to another

As shown in Eq. (31), the squint angle varies with frequency *and* the distance between elements. Another variation of this array is shown in Fig. 43c. As in the previous case, here, too, the phase gradient is realized using true time-delay lines, and the difference between the three layouts is in the implementation. Parallel feed networks (of the type shown in Fig. 44a) are typically used when a nonscanning (with frequency) beam is required. However, they are relatively complex (especially if low sidelobes are required) and therefore occupy significant space on the board. In addition, they radiate, and their spurious radiation can interfere with the radiation of the radiation elements, affecting the sidelobe level and/or the cross-polarization level of the whole array.

When the bandwidth of the array is very small (~1–2%), serial feeds can be used. Serial feeds have been used for decades in slotted waveguide arrays. In the waveguide case, they are of two types: traveling-wave arrays and resonant arrays. The resonant arrays end up in a short, and with the radiating elements separated by

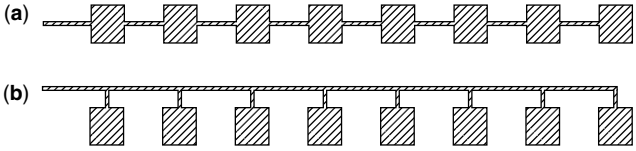


Figure 44. (**a**) two-port serial fed array; (**b**) One-port serial fed array.

a half-guide wavelength, all elements can be fed in phase so that a broadside beam can be obtained.

The traveling-wave arrays are fed at one end and are terminated into a matched load. Since most of the power is radiated through the slots, the matching load absorbs only a small fraction of the incident wave. The spacing between elements is not a half-guide wavelength (to avoid reflection in phase), and the direction of the beam is never broadside.

The main problem with waveguides is that the characteristic impedance of the waveguide cannot be (easily) changed. In the case of microstrip lines, this can be done by simply changing the width of the transmission line. This allows for a greater freedom of design and more types of serially fed microstrip arrays concepts to be introduced. Moreover, the serially fed microstrip array can have any polarization, unlike the slotted waveguide, where the polarization can be linear only.

The wide variety of serially fed microstrip arrays makes it somewhat difficult to divide them into specific groups. One classification would be in arrays where the microstrip element is used as both a two-port device and a one-port device (Fig. 44). In both cases resonant and traveling-wave arrays can be designed. The methods of feeding can vary: microstrip line or aperture-fed (Fig. 45). A full design procedure based on the transmission line model of these arrays is given in Chapter 14 of Ref. 11.

An excellent example of a shaped beam serially fed microstrip array is shown in Fig. 46 [34]. The design is based on the transmission-line model for the patches with measured values for the different components. The widths of the patches and their location are calculated so as to produce the desired radiation pattern.

Figure 47 shows the comparison between the calculated pattern and the measured one, while Fig. 48 shows the change of the radiation pattern with frequency. The one-port microstrip serially fed antennas allow for even greater freedom of design.

Three types of arrays have been described [35]:

1. The first array uses a standard standing-wave feed design. As in Ref. 34, the patch width is varied in order to obtain the desired amplitude taper (Fig. 49a). The transmission line connecting the patch to the main feeder is $\lambda_g/2$ long, so it transforms the input impedance of each patch directly to the main feeder. The characteristic impedance of the main feeder is constant for its entire length.
2. The second design also uses patches of varying widths, but the main feedline is matched at each patch tap point (Fig. 49b).
3. The third array uses a center-fed feed network with each half designed as a traveling-wave array with a main beam angle slightly off broadside. The combination of both halves will yield a broadside beam, which does not *scan* with frequency, but whose shape *will* slightly vary with frequency (Fig. 49c).

Three 16-element arrays with a 22-dB sidelobe level for each of the designs described above were designed and tested (see Table 2 [36]).

In addition to the two classes described above, other types of serially fed microstrip arrays can be mentioned:

1. Linear array with capacitively coupled microstrip patches (Fig. 50)
2. Comb-Line array with microstrip stubs (Fig. 51)

7.3. Planar Arrays

When a narrow pencil beam is required in both planes, planar arrays (rather than linear arrays) have to be used.

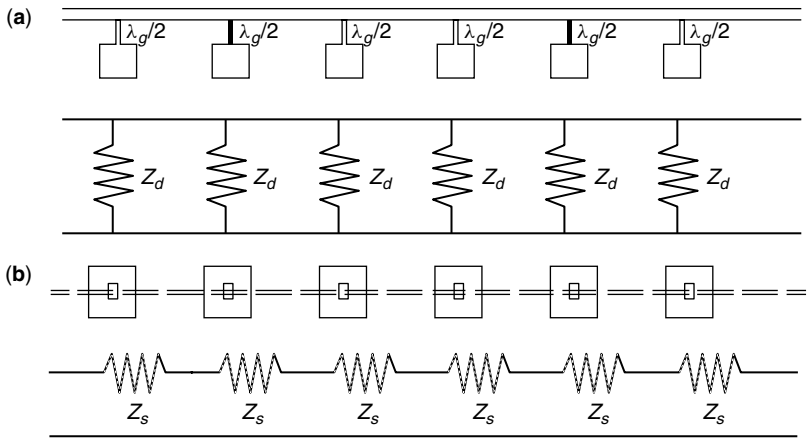


Figure 45. Two-port serial fed microstrip arrays: (**a**) microstrip fed and (**b**) aperture-coupled fed. (Courtesy of Prof. D. M. Pozar, Univ. of Massachusetts at Amherst).

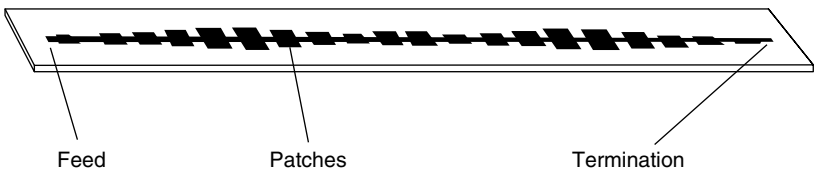


Figure 46. A serial fed microstrip array with a cosec^2 pattern [34].

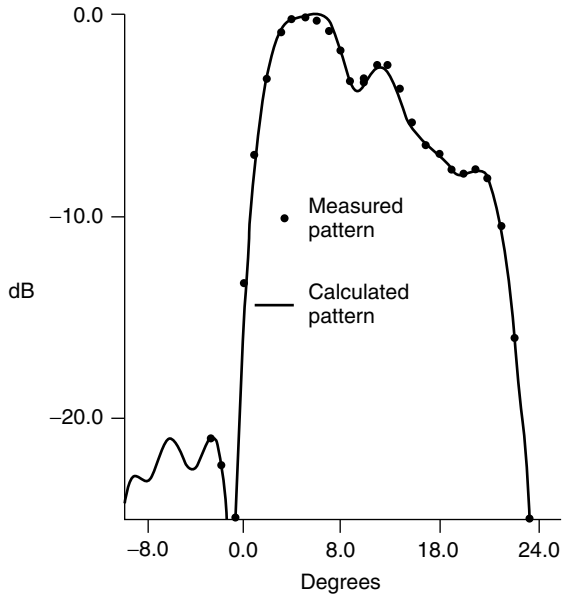


Figure 47. Comparison of measured and calculated amplitude patterns of the cosec^2 patch array [34].

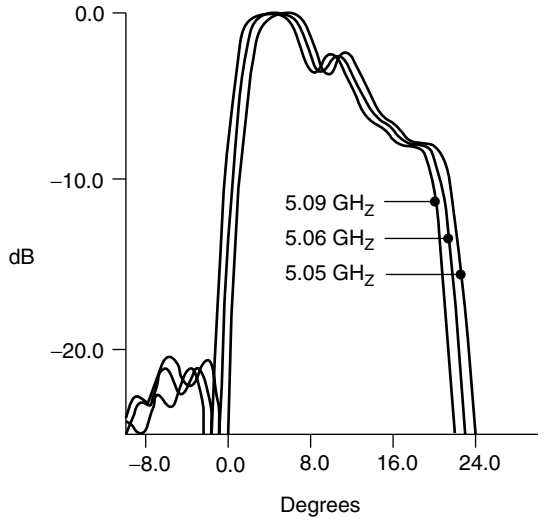


Figure 48. Effect on pattern of a 1% change in frequency [34].

Many configurations have been proposed for planar arrays, in which serial feed and/or parallel feed subarrays can be combined.

The main difficulty in planar arrays is the limited space within the unit cell. To avoid grating lobes, the unit cell has to be no larger than about 0.5–0.8 wavelength, while the typical patch printed on a low dielectric substrate is about 0.3–0.4 wavelength. Therefore, except for arrays with uniform amplitude and phase distribution (where the parallel feed is relatively simple to implement; see Fig. 52), the parallel feed network could be very complex. Such a feed network would couple to the radiating elements, resulting in significant degradation of array performance. In some cases, the degradation of the sidelobe level can be up to 10 dB [41]. To alleviate this problem, a combination between a corporate feed and a parallel feed can be used. Some examples are shown in Fig. 53.

In Fig. 53a square patches are used to yield the same resonance frequency for the two polarizations. For each polarization, four serially fed subarrays are combined by means of a parallel feeding network.

An interesting method to control the sidelobe level has been proposed [40] (Fig. 53b). The power tapering is achieved by connecting the equally wide patches diagonally. Changing the slope of the connecting feeding lines controls the beamwidth. Figure 53c shows a 9×9-element array. Here the serial feed is used in both planes. This is an example of the planar form comb array. The taper required for sidelobes is obtained by simply assigning the appropriate width to the microstrip stubs. An example of interlaced networks is shown Fig. 53d. The antenna consists of two arrays: one operating at 2.45 GHz and the other at 5.8 GHz. The 2.45-GHz radiating element is a rectangular stacked patch with a high aspect ratio. The input impedance of this element is 200 Ω, so a 3D linear transformer was required to reduce the impedance to 100 Ω. Note that the transformer does not change in *width* but in *height* above the ground plane. The length of the transformer was determined to match the bandwidth requirements. The 5.8-GHz elements are suspended square patches. In this design, the interlaced architecture is possible only because of the use of serially fed arrays.

7.4. Scanning Arrays

The array's scanning capability is frequently used in many military as well as commercial applications. This can be done electronically to achieve continuous coverage at very high scan rates. Unlike the situation in mechanical scanning, where the whole antenna is rotated, in electronic scanning, the radiating aperture is fed the appropriate phase distribution, which controls the direction of the main beam. The direction of the main beam is given by Eq. (31).

By scanning the beam of an array, besides the direction of the main beam axis, most of the array

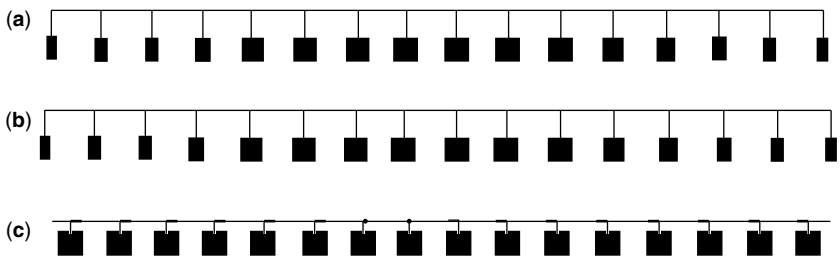


Figure 49. Series fed array designs: (**a**) standing-wave array with element spacing of λ_g; (**b**) traveling-wave array using matched feedlines and an element spacing of λ_g; and (**c**) traveling-wave array with element spacing less than λ_g [35].

Table 2. Summarized Performance of Three Arrays

Array Type/ Parameter	Standing-Wave Array	Matched Traveling-Wave Array	Phase-Compensated Traveling-Wave Array
Impedance BW (calculated)	1.8%	2.0%	4.0%
Impedance BW (measured)	1.7%	1.3%	4.2%
Directivity (calculated)	18.9 dB	18.9 dB	17.9 dB
Gain (calculated)	17.8 dB	17.8 dB	17.3 dB
Gain (measured)	17.4 dB	16.9 dB	16.5 dB
Efficiency (calculated)	77%	78%	88%
Sidelobe level (design)	22 dB	22 dB	20 dB
Sidelobe level (measured)	21 dB	22 dB	21 dB
Pattern BW[a] (measured)	2.3%	2.3%	12%

[a]For sidelobe level remaining below 13 dB.
Source: Courtesy of Prof. David Pozar, University of Massachusetts at Amherst.

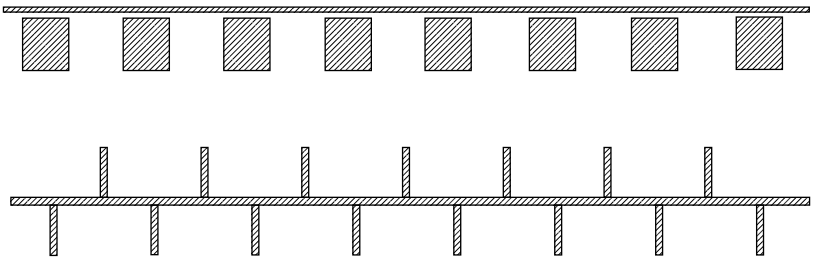


Figure 50. Linear array with capacitively coupled microstrip patches.

Figure 51. Comb-line array with microstrip stubs.

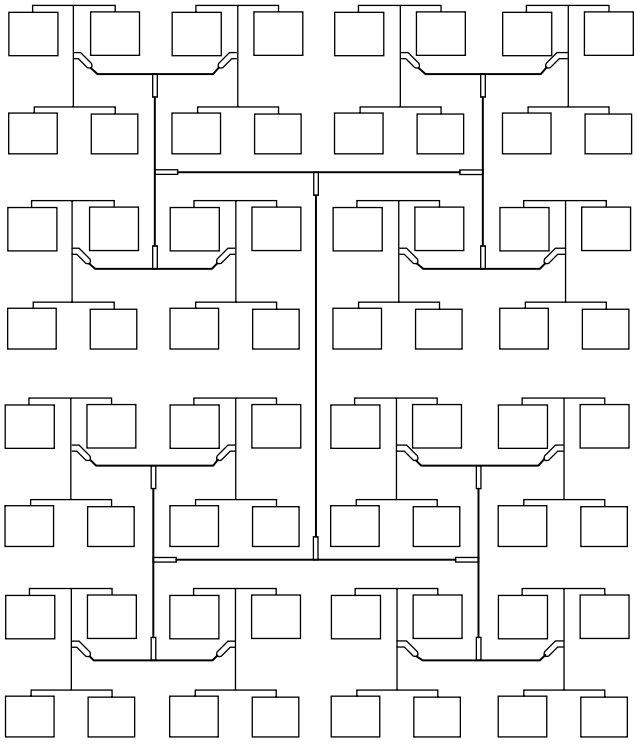


Figure 52. Typical parallel fed microstrip array [38].

characteristics change: beamwidth, radiation pattern, and input impedance. This is because the input impedance and the radiation pattern of each individual element in the array change, as well as the mutual coupling between elements. The *active-element pattern* of an element in an array is defined as the radiation pattern of the array when only that element is driven and all other elements are terminated in matched loads. In the absence of grating lobes, it can be shown that the active-element pattern is given by

$$F(\theta, \phi) = (1 - |R(\theta, \phi)|^2) \cos\theta \tag{33}$$

where

$$R(\theta, \phi) = \frac{Z_{\text{in}}(\theta, \phi) - Z_{\text{in}}(0, 0)}{Z_{\text{in}}(\theta, \phi) + Z_{\text{in}}(0, 0)} \tag{34}$$

and $|R(\theta, \phi)|$ is the active reflection coefficient. The term, $Z_{\text{in}}(0, 0)$ is the input impedance when the beam is at broadside and the array is assumed to be matched for maximum array gain.

Depending on the array geometry and its *physical implementation*, for some scanning angles, the active reflection coefficient might be close to unity. In this case, no power is radiated by the array. This phenomenon is generally known as *scan-blindness*. It was initially studied for the infinite array case; however, the same theory is applicable for finite arrays [43, 47]

The following example deals with a 9×9 array of rectangular patches printed on a $0.06\lambda_0$ thick substrate with a relative dielectric constant $\varepsilon_r = 12.8$ [45]. The calculations of the different infinite array parameters are compared to the infinite array case. Figure 54 shows the geometry of the array, and Fig. 55 summarizes the results.

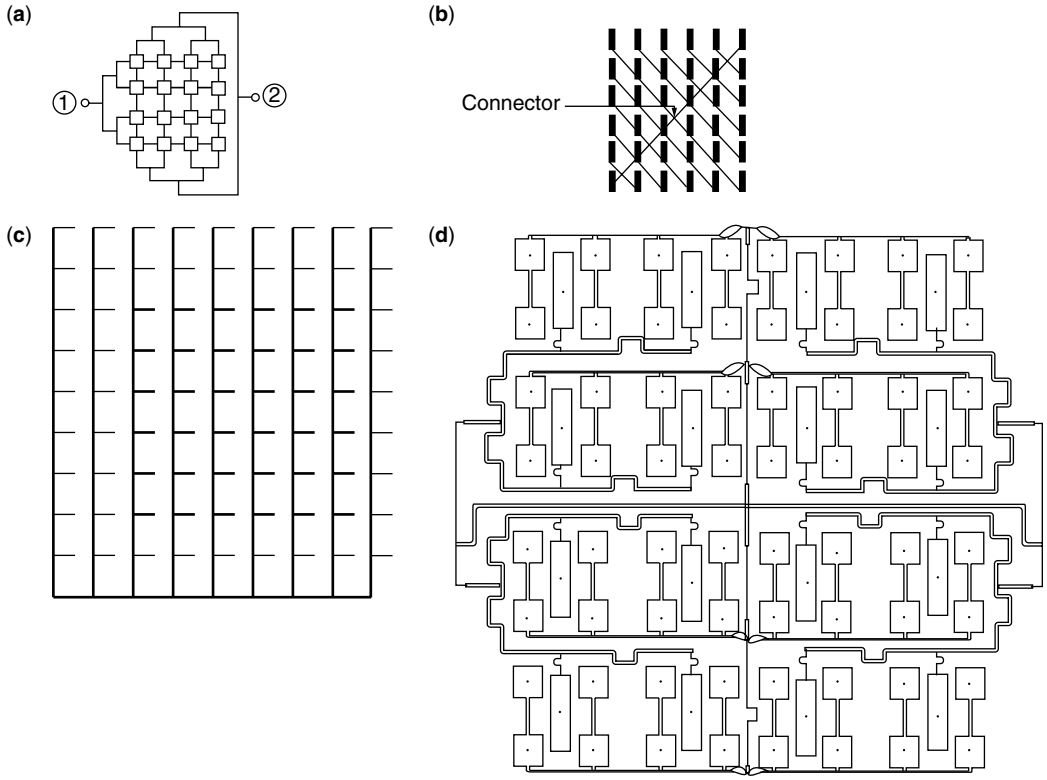


Figure 53. Examples of microstrip planar arrays: (**a**) a dual-polarized 4 × 4-element microstrip array (port 1 is for horizontal polarization; port 2 is for vertical polarization [39]); (**b**) a Cross-fed array [40]; (**c**) J-band planar array of nine linear arrays with nine cophase stubs [41]; and (**d**) a dual-band (2.45/5.7-GHz) planar array [42].

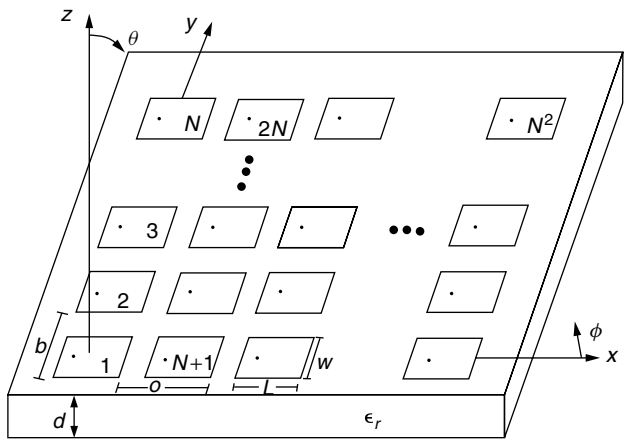


Figure 54. Geometry of the finite array of rectangular microstrip patches [45].

BIOGRAPHY

Naftali Herscovici was born in Bucuresti, Romania in 1954. He received his B.Sc. and M.Sc. from the Technion, Haifa, Israel and his Ph.D. from the University of Massachusetts, Amherst, in 1978, 1985, and 1992 respectively. Between 1982 and 1989 he was employed by Rafael, Haifa, Israel as an Antenna Research Engineer; there he was engaged in research and development of microwave antennas. He is currently the President and Founder of Anteg, Inc., Framingham, Massachusetts. His research interests include microstrip antennas and arrays, reflector antennas and feeds, pattern synthesis, and antenna modeling. Dr. Herscovici is the author of over 50 technical papers in various journal and conference publications.

BIBLIOGRAPHY

1. G. A. Deschamps, Microstrip microwave antennas, *Proc. 3rd USAF Symp. Antennas*, 1953.
2. H. Gutton, and G. Baissinot, Flat aerial for ultra high frequencies, French Patent 70313 (1955).
3. E. V. Byron, A new flush-mounted antenna element for phased array application, *Proc. Phased Array Antenna Symp.*, 1970, pp. 187–192.
4. R. Munson, Conformal microstrip antennas and microstrip phased arrays, *IEEE Trans. Antennas Propag.* **AP-22**: 74–78 (1974).
5. E. H. Newman, and P. Tulyathan, Microstrip analysis technique, *Proc. Workshop Printed Circuit Antennas*, New Mexico State Univ., Oct. 1979, pp. 9.1–9.8.
6. I. J. Bahl, Build microstrip antennas with paper-thin dimensions, *Microwaves* **18**: 50–63 (Oct. 1979).
7. K. R. Carver, Practical analytical techniques for the microstrip antenna, *Proc. Workshop Printed Circuit Antennas*, New Mexico State Univ., Oct. 1979, pp. 7.1–7.20.

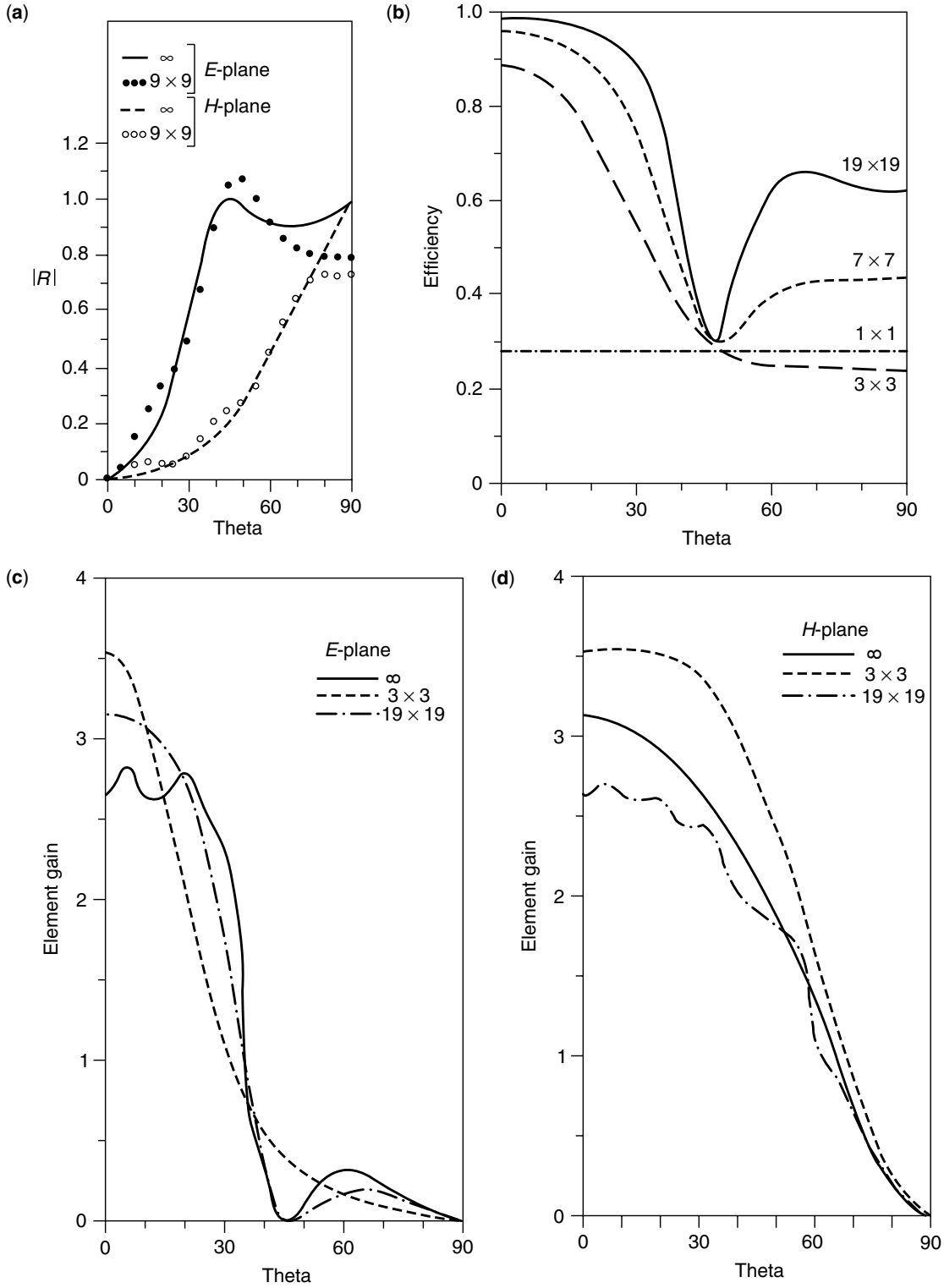


Figure 55. Calculated results for finite patch arrays on $0.06\lambda_0$ thick substrate with a relative dielectric constant $\varepsilon_r = 12.8$ ($a = b = 0.5\lambda_0$, $L = 0.1074\lambda_0$, $W = 0.15\lambda_0$, $X_p = -L/2$, $Y_p = 0$) [45]: (**a**) Reflection coefficient magnitude versus scan angle (E and H planes) for a finite (9×9, center element) patch array, compared with infinite array results; (**b**) efficiency of a finite patch array versus E-plane scan angle for various sizes; (**c**) E-plane active-center-element gains for patch arrays of various sizes; (**d**) H-plane active-center-element gains for patch arrays for various sizes, (**e**) E-plane active-element gain patterns for various elements of a 13×13 patch array; and (**f**) H-plane active-element gain patterns for various elements of a 13×13 patch array.

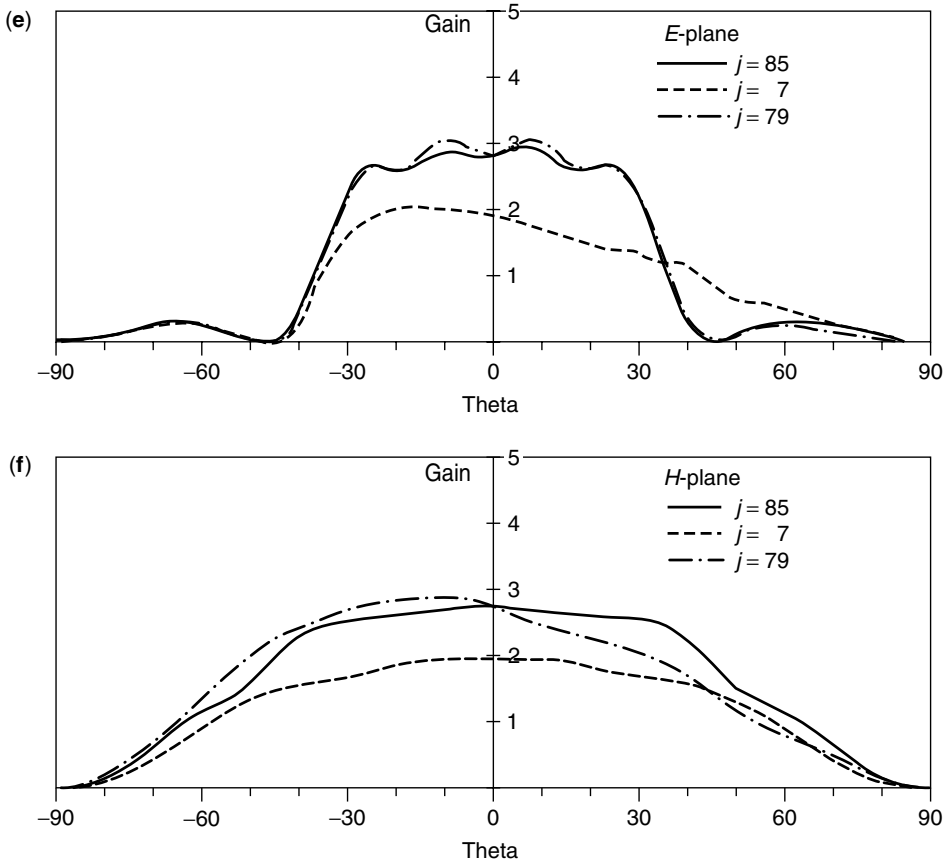


Figure 55. (*Continued*)

8. I. J. Bahl and P. Bhartia, *Microstrip Antennas*, Artech House, 1980.
9. W. F. Richard, Y. T. Lo, and D. D. Harrison, Theory and experiment on microstrip antennas, *Electron. Lett.* **15**: 42–44 (1979).
10. A. G. Derneryd and A. G. Lind, Extended analysis of rectangular microstrip antennas, *IEEE Trans. Antennas Propag.* **AP-27**: 846–849 (1979).
11. J. R. James and P. S. Hall, *Handbook of Microstrip Antennas*, Peter Peregrinus, London, 1989.
12. R. Garg, P. Bhartia, I. Bahl, and A. Ittipiboon, *Microstrip Antenna Design Handbook*, Artech House, Norwood, MA, 2001.
13. N. Herscovici, Z. Sipus, and D. Bonefacic, Circularly polarized single-fed wide-band microstrip elements and arrays, *1999 IEEE Int. Antennas Propag. Symp. Dig.* **37**: 280–283 (June 1999).
14. J. Huang, A technique for an array to generate circular polarization with linearly polarized elements, *IEEE Trans. Antennas Propagat.* **34**: 1113–1124 (Sept. 1986).
15. M. Haneishi, S. Yoshida, and N. Goto, A broadband microstrip array composed of single-feed type circularly polarized microstrip antennas, *1982 IEEE Int. Antennas Propag. Symp. Dig.* **20**: 160–163 (May 1982).
16. T. Teshirogi, M. Tanaka, and W. Chujo, Wideband circularly polarized array antenna with sequential rotations and phase shifts of elements, *Proc. Int. Symp. Antennas and Propagation*, Japan, 1985, pp. 117–120.
17. D. M. Pozar, Considerations for millimeter wave printed antennas, *IEEE Trans. Antennas Propag.* **31**: 740–747 (Sept. 1983).
18. N. Herscovici, A wide-band single-layer patch antenna, *IEEE Trans. Antennas Propag.* **46**: 471–474 (April 1998).
19. R. Zetner, J. Bartolic, and E. Zetner, Electromagnetically coupled butterfly patch antenna, *J. Int. Nice Antennes* 588–591 (Nov. 1996).
20. H. F. Pues and A. R. Van de Capelle, An impedance-matching technique for increasing the bandwidth of microstrip antennas, *IEEE Trans. Antennas Propag.* **37**: 1345–1354 (Nov. 1989).
21. S. M. Duffy, An enhanced bandwidth design technique for electromagnetically coupled microstrip antennas, *IEEE Trans. Antennas Propag.* **48**: 161–164 (Feb. 2000).
22. R. Kastner, E. Heyman, and A. Sabban, Spectral domain iterative analysis of single- and double-layered microstrip antennas using the conjugate gradient algorithm, *IEEE Trans. Antennas Propag.* **36**: 1204–1212 (Sept. 1988).
23. F. Croq and D. M. Pozar, Millimeter-wave design of wide-band aperture-coupled stacked microstrip antennas, *IEEE Trans. Antennas Propag.* **39**: 1770–1776 (Dec. 1991).
24. A. Sabban, A new broadband stacked two-layer microstrip antenna, *1983 IEEE Int. Antennas Propag. Symp. Dig.* **21**: 63–66 (May 1983).
25. G. Kumar and K. C. Gupta, Broadband microstrip antennas using coupled resonators, *1983 IEEE Int. Antennas Propag. Symp. Dig.* **21**: 67–70 (May 1983).

26. G. Kumar and K. C. Gupta, Nonradiating edges and four edges gap-coupled multiple resonator broad-band microstrip antennas, *IEEE Trans. Antennas Propag.* **33**: 173–178 (Feb. 1985).

27. G. Kumar and K. C. Gupta, Directly coupled multiple resonator wide-band microstrip antennas, *IEEE Trans. Antennas Propag.* **33**: 588–593 (June 1985).

28. F. Croq and D. M. Pozar, Multifrequency operation of microstrip antennas using aperture-coupled parallel resonators, *IEEE Trans. Antennas Propag.* **40**: 1367–1374 (Nov. 1992).

29. S. D. Targonski, R. B. Waterhouse, and D. M. Pozar, An aperture coupled stacked patch antenna with 50% bandwidth, *IEEE Antennas Propagation Symp. Dig.*, Baltimore, MD, July 1996, 18–22.

30. S. D. Targonski, R. B. Waterhouse, and D. M. Pozar, Design of wide-band aperture-stacked patch microstrip antennas, *IEEE Trans. Antennas Propag.* **46**: 1245–1251 (Sept. 1998).

31. S. D. Targonski and R. B. Waterhouse, Reflector elements for aperture and aperture coupled microstrip antennas, *1997 IEEE Int. Antennas Propag. Symp. Dig.* **35**: 1840–1843 (June 1997).

32. D. M. Pozar, Input impedance and mutual coupling of rectangular microstrip antennas, *IEEE Trans. Antennas Propag.* **30**: 1191–1196 (Nov. 1982).

33. R. P. Jedlicka, M. T. Poe, and K. R. Carver, Measured mutual coupling between microstrip antennas, *IEEE Trans. Antennas Propag.* **29**: 147–149 (Jan. 1981).

34. B. B. Jones, F. Y. M. Chow, and A. W. Seeto, The synthesis of shaped patterns with series-fed microstrip patch arrays, *IEEE Trans. Antennas Propag.* **30**: 1206–1212 (Nov. 1982).

35. D. M. Pozar and D. H. Schaubert, Comparison of three series fed microstrip array geometries, *1993 IEEE Int. Antennas Propag. Symp. Dig.* **31**: 728–731 (June 1993).

36. D. M. Pozar, private communication.

37. P. S. Hall and C. M. Hall, Coplanar corporate feed effects in microstrip patch array design, *IEE Proc. Part H*, vol. 135, June 1988 pp. 180–186.

38. D. M. Pozar, *Workshop of Antennas for Wireless Communications*, Nov. 1998.

39. A. G. Derneryd, Microstrip array antenna, *Proc. 6th European Microwave Conf.*, 1976, 339–343.

40. J. C. Williams, Cross fed printed aerials, *Proc. 7th European Microwave Conf.*, 1977, 292–296.

41. J. R. James and P. S. Hall, Microstrip antennas and arrays, Part 2—New array design technique, *IEEE J. Microwaves Opt. Acoust.* **1**: 175–181 (1977).

42. N. Herscovici, New considerations in the design of microstrip antennas, *IEEE Trans. Antennas Propag.* **46**: 807–812 (June 1998).

43. R. C. Hansen, *Microwave Scanning Arrays,* Vol. 2, Academic Press, 1966.

44. R. J. Mailloux, *Phased Array Antenna Handbook*, Artech house, 1994.

45. D. M. Pozar, Finite phased arrays of rectangular microstrip patches, *IEEE Trans. Antennas Propag.* **34**: 658–665 (May 1986).

46. D. M. Pozar and D. H. Schaubert, Analysis of an infinite array of rectangular microstrip patches with idealized probe feeds, *IEEE Trans. Antennas Propag.* **32**: 1101–1107 (Oct. 1984).

47. D. M. Pozar and D. H. Schaubert, Scan blindness in infinite phased arrays of printed dipoles, *IEEE Trans. Antennas Propag.* **32**: 602–610 (June 1984).

MICROSTRIP PATCH ARRAYS

R. B. WATERHOUSE
K. GHORBANI
RMIT University
Melbourne, Australia

1. INTRODUCTION

Microstrip patch antennas have long been touted as one of the most versatile radiating structures. These printed radiating elements have several well-known advantages over conventionally styled antennas based on wires and metallic apertures including their low profile, low cost, robustness, and ease of integration with other components. Since 1980 or so this once considered problematic antenna has matured into one of the most commonly used interfaces for free-space/wireless communications. Most mobile communication base stations and handset terminals as well as spaceborne communication systems incorporate this form of radiator.

As a single radiating element, the microstrip patch antenna is generally classified as a low–moderate-gain antenna with gains in the order of 5–8 dBi in its conventional form. One critical advantage of the microstrip patch over its counterparts, which is related to some of the features mentioned before, is the relative ease in which these structures can be integrated or combined to form an array of antennas. By doing so greatly increases the flexibility in shaping the radiation pattern and other features of the antenna, which is consistent with arraying wire, metallic and other forms of radiators. However the distinct advantages of arraying microstrip elements are the ease in fabricating the entire structure, the simplicity of the array layout as well as the low cost of production. The fabrication of these antennas is based on printed circuit board (PCB) etching processes that has minimal labor costs.

In this article we review the development of arrays based on microstrip patch technology. Firstly we discuss the fundamental styles of linear arrays that can be developed using microstrip patches, namely, series feed, corporate feed and a combination feed technique. For each of these methods advantages, issues and design cases are given. The scanning performance (or radiation control) of a linear array is discussed and a design case is once again given. The concepts introduced for linear arrays are then expanded on to investigate planar arrays and methods on how these radiating structures can be developed are presented. Some printed antenna alternatives are summarized that can yield high-gain solutions, with minimal complexity. These printed antennas can overcome the feed loss problems associated with very large planar

arrays. Finally the scanning performance of large planar arrays of microstrip patch antennas is examined and once again the parameters affecting the control of the radiation distribution as well as the limiting performance are discussed.

2. LINEAR ARRAYS

Examining linear arrays of any antenna is probably the easiest means to see how the radiation performance can be controlled in a particular direction or dimension. Of course, the concepts developed or derived from a linear array can be readily expanded to a planar, or a two-dimensional solution. There are numerous books and articles on array theory and the reader should consult these to understand the fundamental properties of arrays [e.g., 1]. There are three types of microstrip patch linear arrays: the series-fed configuration, the corporate (or parallel)-fed geometry, and the combination technique. These methods are examined herein.

2.1. Series-Fed Arrays

One of the first realizations of a microstrip patch array was the series-fed array [e.g., 2]. Here each element of the array is connected in series via an arrangement of transmission lines. Figure 1 shows a schematic diagram of an 8-element series-fed array consisting of edge-fed patches. The array is fed from the left and is classified as a standing-wave array. Series-fed arrays have been developed in waveguide realizations for decades; however, microstrip forms have much more flexibility. This is due mainly to the fact that it is easy to change the impedance of the microstrip feedlines between the radiating elements to give the desired amplitude taper [3]. The patch width can also be varied to give this same effect.

The advantages of microstrip patch arrays utilizing a series feed configuration over other forms (to be discussed later) include it having a simple, more compact feed network, as evident from Fig. 1, and lower feedline loss. However, this form of microstrip patch array does suffer from several drawbacks. The most fundamental issue is the narrow radiation bandwidth of the array, which is typically much narrower than the inherent impedance bandwidth of the individual microstrip elements. There are only several reported cases of series-fed microstrip arrays in the literature, and these have bandwidths typically only fractions of a percent. As microstrip patches in their original form have a high Q value, placing them in series means that each will have a direct impact on the other, and therefore if there are any errors in fabrication or factors not taken into consideration with the design (such as mutual coupling), the overall array performance will be degraded. Because the power to be supplied to each element must by transferred from the previous element (see Fig. 1), the rapid impedance variation of the conventional microstrip patch inherently hinders the delivery of the power to the other elements. Although there have been several techniques over the years to increase the impedance bandwidth of individual microstrip patches, such as a proximity coupled or aperture-coupled patch, incorporating a series feed array solution of these radiators removes the open- or short-circuited tuning stub, reducing the number of degrees of freedom of these non-contact-excitation methods and hence their flexibility.

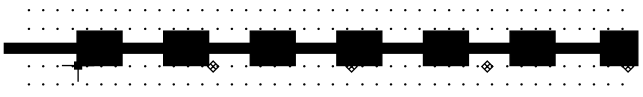


Figure 1. Schematic diagram of 8-element series-fed linear array of microstrip patches.

Figure 2 shows a 4-element edge-fed series microstrip array. The parameters for the four elements are shown in Fig. 2. The length of elements and distance between them were varied to achieve maximum gain near broadside using a full-wave simulator. The feedlines between the elements are 100-Ω transmission lines, so the disturbance in the field of radiators can be minimized. The overall antenna is matched to 50 Ω by a quarter-wave transformer. The widths of antenna elements were varied to reduce the sidelobe levels. The return loss response and the E-plane radiation performance of the array across the 10-dB return loss bandwidth of 1% is shown in Fig. 3a,b, respectively. As can be seen from Fig. 3b, the radiation pattern remains generally constant across the matched impedance bandwidth, unlike some of the early cases of series-fed microstrip patch arrays. There is a slight asymmetry in the radiation pattern that is due to the presence of the feed network. The H-plane pattern is similar to a conventional edge-fed microstrip patch [e.g., 1]. The cross-polarization level in both principal planes (E and H planes) was less than 40 dB below the copolar fields. The relatively constant radiation performance of the antenna can be attributed to using a full-wave simulator to synthesize the antenna, software tools that were not available in the early days of microstrip patch technology development or were simply too slow for

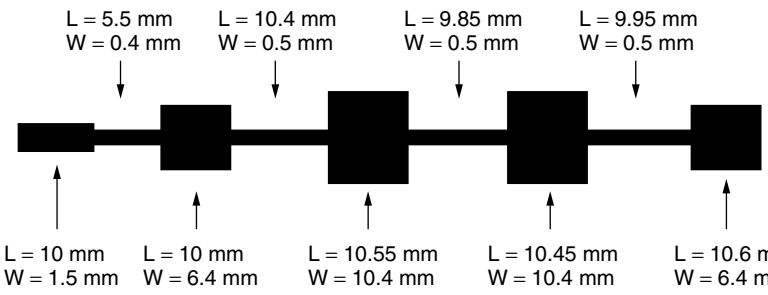


Figure 2. Schematic diagram of 4-element series-fed linear array of microstrip patches.

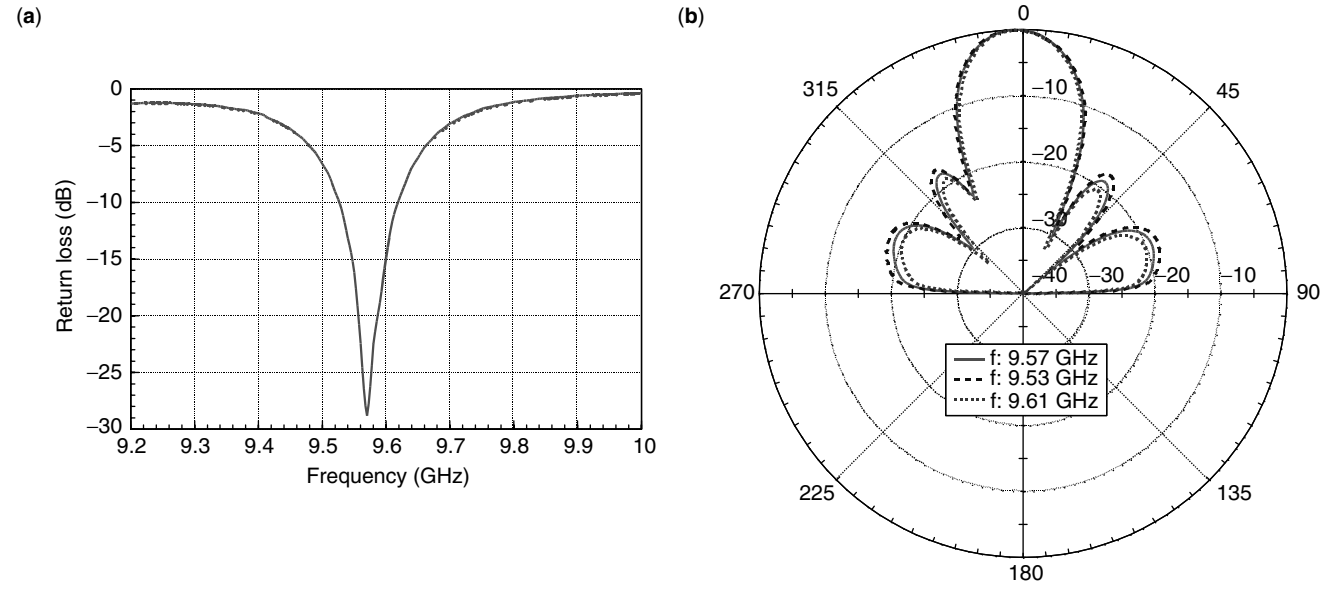


Figure 3. Characteristics of 4-element series fed array: (**a**) return loss; (**b**) radiation patterns.

design purposes. Utilizing such tools as well as enhanced bandwidth elements (such as stacked patches) should see an improvement in the performance of series-fed arrays, but probably not to the same degree as corporate (or parallel) fed microstrip arrays.

2.2. Parallel Fed Arrays

Parallel or corporate fed microstrip patch arrays are the most common type of array using microstrip patch technology. Here, unlike the series-fed array, each element has its own excitation transmission line, which can be made independent of the feedlines of the other elements as well as the other elements of the array. Figure 4 shows a schematic diagram of an 8-element corporate feed array of edge-fed microstrip patches. As can be seen from the figure, each element has its own excitation transmission line. Each of these transmission lines is then connected together via a series of two-way power combiners, although three-way dividers are commonly used if an odd number of elements are used in the array. The power combiners can either be reactive, such as shown in Fig. 4, or based on Wilkinson dividers. The Wilkinson divider gives broader band isolation between the elements at the expense of increased complexity and also loss. It should be noted that most microstrip patches have impedance bandwidths smaller than that of a reactive power divider.

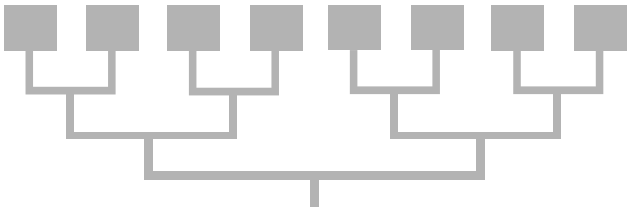


Figure 4. Schematic diagram of an 8-element corporate fed linear array of microstrip patches.

Of all the array formats, parallel configurations have the broadest bandwidths, in some cases even greater than that of the individual elements of the array. This effect can be attributed to the cancellation of unwanted reflections of power within the feed network. The good isolation between the individual feedlines allows the ready incorporation of phase shifters to allow scanning of the radiation beam of the array (referred to later in this section) as well as amplitude tapers to reduce the sidelobe level. An excellent paper outlining how to do this and the possible source of error is available [4]. The good isolation of the parallel feed allows the designer to separately address the issues related to the individual microstrip patch (the basis of the array) and then the feed network. Such an approach significantly reduces the computational power required to successfully design the array. Because of all these features, corporate fed microstrip patch arrays are utilized in many applications such as mobile base-station antennas.

Figure 5 shows a schematic diagram of an 8-element corporate array of aperture-coupled microstrip patches. For details on the design of aperture-coupled patches

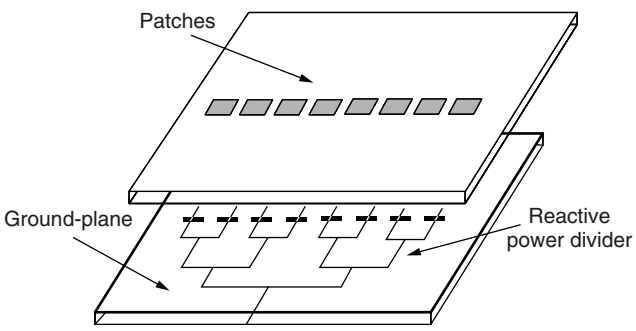


Figure 5. Schematic diagram of 8-element corporate fed linear array of aperture coupled microstrip patches.

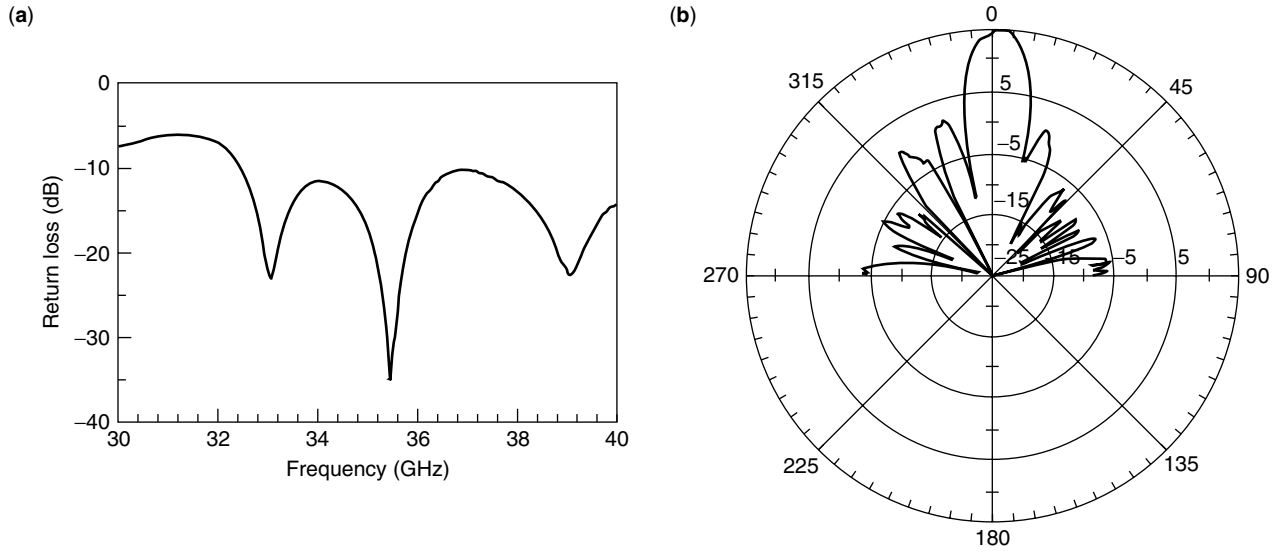


Figure 6. Characteristics of 8 element corporate fed array of aperture coupled patches: (**a**) measured return loss; (**b**) *H*-plane radiation pattern.

please refer to the article by Targonski and Pozar [5]. The array utilizes seven reactive power dividers to feed the elements. As mentioned before, reactive power dividers are more efficient than their counterparts. The array was designed for fiber radio applications at millimeter-wave frequencies [6]. The measured return loss of the array is given in Fig. 6a, and the 10-dB return loss bandwidth is approximately 30%. A sample of the *H*-plane radiation pattern is shown in Fig. 6b. The *E*-plane pattern, not shown here, is similar to a conventional microstrip patch element. The array has a gain of 15 dBi across the entire 10-dB return loss bandwidth, and the cross-polarization levels were less than 20 dB below the copolar fields in both principal planes.

2.3. Combination Feed Arrays

There is a third class of microstrip array, which is a combination of the series and parallel feed methods. Here a common feedline is used for the entire array and each element taps power off this feedline. To ensure that the bandwidth is greater than the conventional series-fed array, each tap and section of the common feedline is impedance matched. Thus the whole design of the array simplifies itself to uncomplicated impedance matching to ensure good impedance bandwidth as well as the appropriate distribution of power. Figure 7 shows a schematic diagram of how an 8-element version of this array can be realized. First, the array is split into two symmetrical parts, one of which is shown in Fig. 7. Here each element is designed for an input impedance at resonance of 200 Ω. Doing so allows for a relatively straightforward impedance-matching design to ensure that equal power is distributed to each microstrip patch. The feedlines connected to each element are also made to have a characteristic impedance of 200 Ω and a length of $\lambda_g/2$, where λ_g is the guided wavelength in microstrip. Doing this minimizes the effects of error in the design of the microstrip patch, which will further impact the performance of the array. It is also difficult to fabricate 200-Ω transmission lines on some material, so the $\lambda_g/2$ length transforms the input impedance of the patch to the central feedline (refer to Fig. 7).

An 8-element combination array was designed and developed centered at 9 GHz. A photograph of the array is shown in Fig. 8 (see also Fig. 9). The impedance bandwidth was measured as 5%. An example of the radiation patterns in the *H* and *E* planes of the antenna are shown in Fig. 9a,b, respectively. The *H*-plane pattern shows the expected focusing of the beam in the plane of the array (note that the slight off-broadside pattern is due to alignment errors in the measurement setup). A small

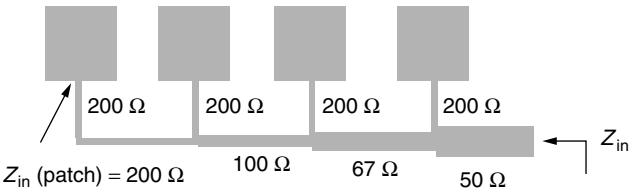


Figure 7. Schematic diagram outlining design of combination feed linear array.

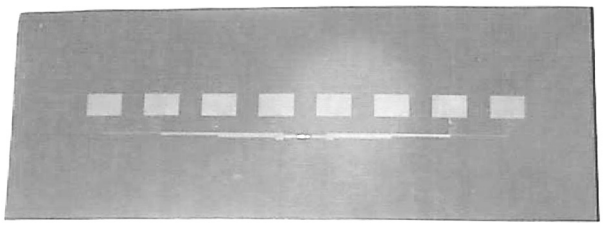


Figure 8. Photograph of 8-element combination feed linear array.

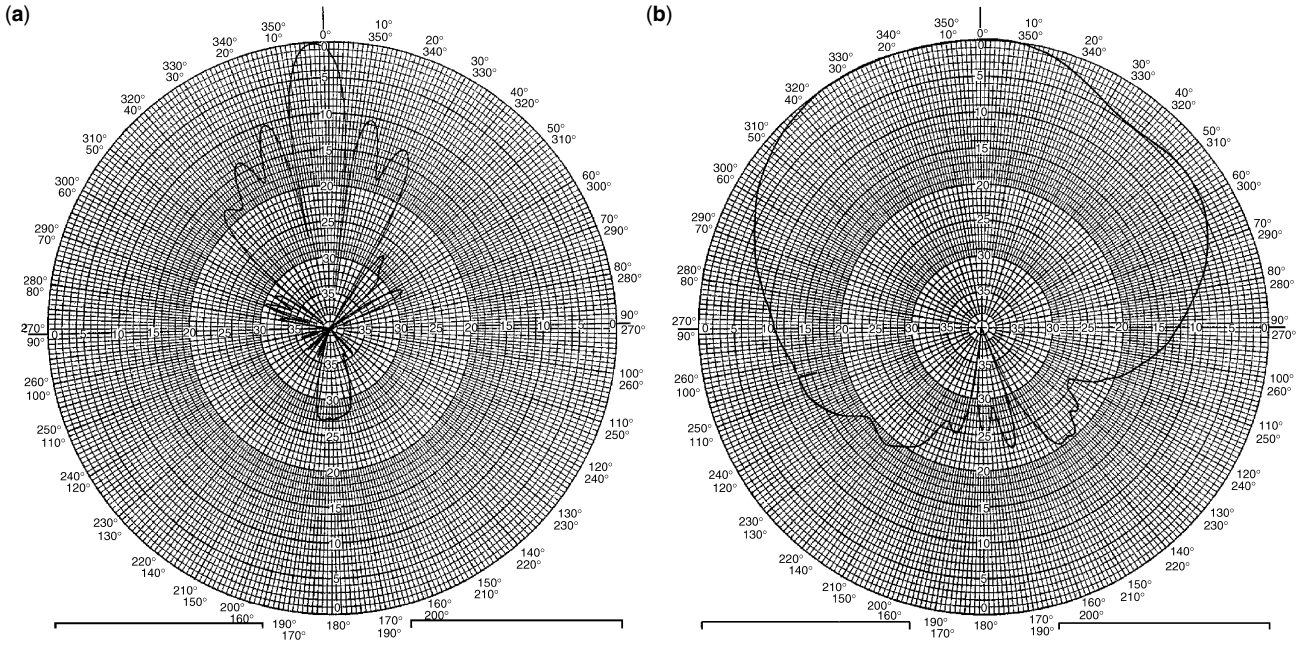


Figure 9. Radiation performance of 8-element combination feed linear array: (**a**) *H* plane; (**b**) *E* plane.

scalping of the pattern is evident in the *E* plane. This is due to the radiation from the feed network. The gain of the array was measured as 15 dBi across the impedance bandwidth. The overall performance of a combination feed array lies somewhere between that of the series-fed array and the parallel array. Its impedance and radiation bandwidth (3 dB gain) is greater than the series array but less than that of the corporate array. Its efficiency is greater than the parallel configuration, although less than the series method.

2.4. Scanned Linear Arrays

The linear array designs considered above are fixed-beam examples, with the main beam directed toward broadside. It is relatively straightforward to point this beam at a fixed angle off broadside by simply inserting a constant phase between the elements. There are many applications. For example, satellite communications, which require a beam that can be scanned or continually steered to ensure contact between a moving object (say, an aircraft) and a stationary or a moving object (say, a constellation of satellites) can be maintained at all times. Because microstrip patches can readily be formed into arrays and easily connected to phase shifting circuitry, these radiators are prime candidates for most phased-array applications. Microstrip elements have broad radiation patterns, which means that arrays of these elements should be able to be scanned to large angles, approaching endfire. This issue will be discussed in Section 3.

Of the three configurations presented, corporate feed arrays are the easiest to control the phasing to its elements. Figure 10 shows a photograph of a linear phased-array based on a corporate feed arrangement mounted on a test jig. The 6.4 GHz 8-element array is located at the top of the photograph, with the radome (a layer of Duroid 5880) that covers the top patches evident in the photograph (the white-grayish rectangular region). The size of the ground plane of the array is 22×10 cm. The array consists of edge-fed stacked patches fed by a 90° branchline coupler to produce circular polarization, a common requirement for satellite communications and

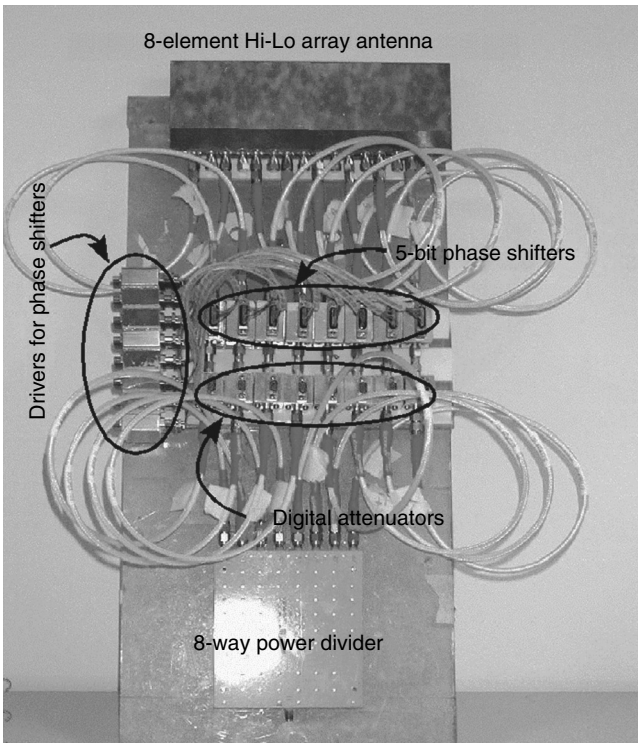


Figure 10. Photograph of linear 8-element phased array of microstrip patches.

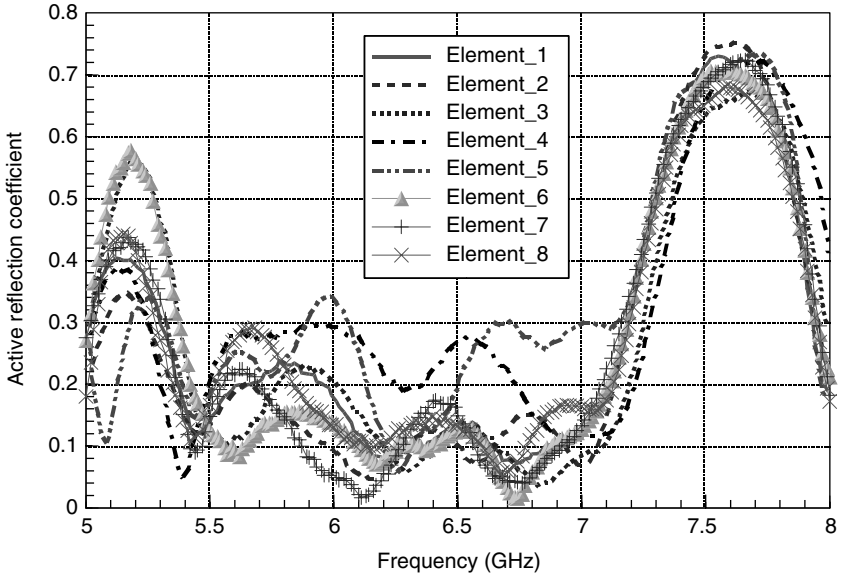


Figure 11. Active reflection coefficient of 8-element linear phased array.

easily achievable using microstrip technology. Feeding the left-hand-side port while terminating the right-hand-side port with matched loads results in right-hand circular polarization (RHCP) generation. Reversing the feed and terminated port results in left-hand circular polarization (LHCP) generation. For the experiments conducted here, only the RHCP was investigated.

Eight phase-matched cables are used to connect the stacked patch array to a bank of 5-bit switched-line phase shifters. As the losses of the phase shifters vary at different phase settings, a bank of 3-bit digital attenuators is included to achieve amplitude balance. In this experiment, the maximum phase error is $\pm 8^\circ$ and the maximum amplitude error is ± 0.7 dB.

Figure 11 shows the measured active reflection coefficient [7] of each element in the array at broadside with the other elements of the array terminated with matched loads. This is a common measurement procedure to ascertain the performance of the array. The worst-case measured 10 dB return loss for each element was 28.6%, centered at 6.3 GHz. The significantly increased impedance bandwidth, compared to the predicted individual stacked patch case of 20% can be attributed to the feed network. Such feed-networks typically cancel unwanted reflections as mentioned previously. It is interesting to note that most of the active reflection coefficients for the 8 elements have similar responses between 5.3 and 7.5 GHz, with the exception of elements 4 and 5. Although these elements still satisfy the 10-dB return loss criteria over the same bandwidth as the other elements, their active reflection coefficients are marginally higher. This may be due to a soldering problem where the microstrip lines are connected to the appropriate SMA-style connectors.

The radiation patterns and axial ratio of the scannable printed array were measured at a variety of frequencies and scan angles. A sample of the results is presented in Fig. 12. The array can be readily scanned to $\pm 45^\circ$ while maintaining an axial ratio of less than 3 dB. The gain across the scanned range of angles and frequencies is approximately 3 dB lower than the expected or predicted values due to the insertion loss of the phasing module.

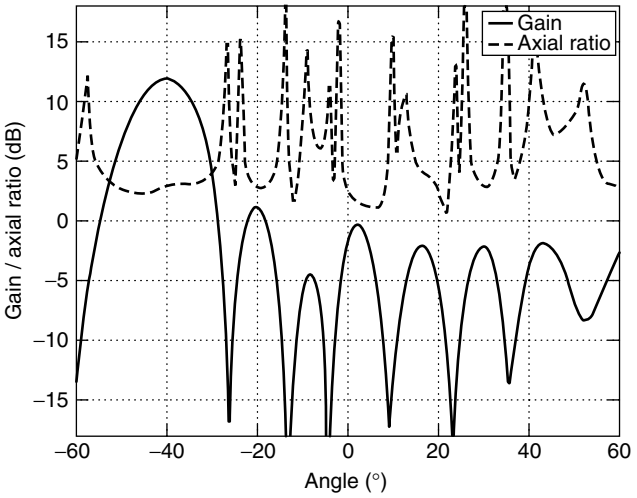


Figure 12. Sample of radiation performance of linear phased array.

3. PLANAR ARRAYS

Having outlined the fundamental structures for linear arrays in the previous section, it is relatively straightforward to extend these configurations into planar, or two-dimensional arrays. Thus you can effectively have a planar series-fed, corporate fed or a combination feed array. However, it is probably easier to categorize planar arrays in two styles: fixed-beam or scanned. Fixed-beam can consist of any of the three linear arrays introduced. Scanned arrays tend to always consist of corporate style feeding, simply because it is the easiest to achieve independent phase (and amplitude) control of the elements of the array. In this section we examine fixed and scanned beam planar arrays.

3.1. Fixed-Beam Planar Arrays

Looking through the literature, to the authors' knowledge there does not appear to be any case of planar series-fed arrays. This is intuitive, for as the length of the array increases, or in the planar case, the area, the elements at the end of the array are less likely to receive any power from the source. Thus these elements in a series feed array are redundant. It is perhaps possible to develop small (say, 8-element) two-dimensional arrays of series-fed microstrip patch elements; however, the shortcomings highlighted previously for the series-fed linear array would still hold for this case. There is reported in the literature a case of several linear series arrays connected in parallel, using the impedance-matching procedure summarized earlier [8].

It is possible to create a planar version of the combination feed array. Figure 13 shows a photograph of a 32-element array. The microstrip array consists of combining four of the 8-element linear arrays considered in the previous section. Once again, to combine these arrays requires the use of impedance matching and quarter-wave transformers. The radiation patterns of the array in both the E and H planes are shown in Fig. 14. The focusing of the radiation toward broadside is evident in this figure. The gain of this array was measured as 21 dBi and the impedance bandwidth as 5%. In Fig. 13 the $\lambda_g/2$ lines that feed the elements of each linear array have been folded on themselves to ensure that the array spacing in the E plane is not too large. The array spacing in each plane is $0.8\ \lambda_0$ to ensure maximum directivity [1].

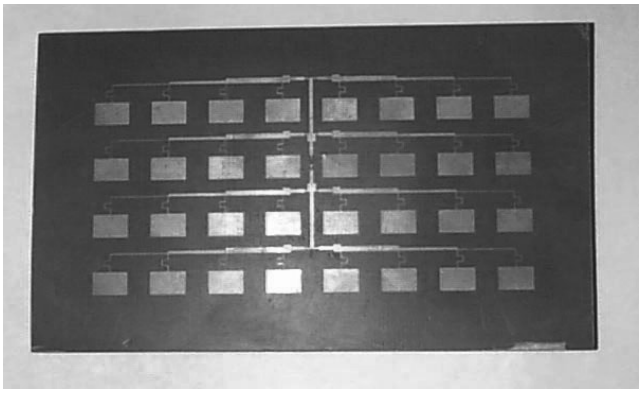


Figure 13. Photograph of 32-element planar combination array.

By far the most common type of fixed-beam microstrip patch array is based on the corporate feed [9]. These planar arrays are utilized in applications such as millimeter wave collision avoidance radar for vehicles, local multipoint distribution services and imaging. A schematic diagram of a 256-element corporate fed patch array is shown in Fig. 15 [10]. The design of these arrays can be somewhat complicated, not so much in terms of the antenna element design, but because of the feed network layout. A good rule of thumb to minimize spurious radiation from feedlines is to keep the structure as symmetric as possible, which tends to minimize cross-polarization levels and to use thin transmission lines. Levine et al. have contributed an excellent paper on the effect of the feed network on the overall performance of a corporate fed microstrip patch array [10]. In this paper, it was shown that as the array gets larger, the loss associated with the feed network gets more and more until it can be substantial. For a 32×32-element array, the loss was more than 7.5 dB. Table 1 shows a comparison of planar arrays of microstrip patches versus reflectors with efficiencies of 50% [10]. The table highlights the issues related to large patch arrays. We can see from this table that although the directivity increases as the number of elements increase

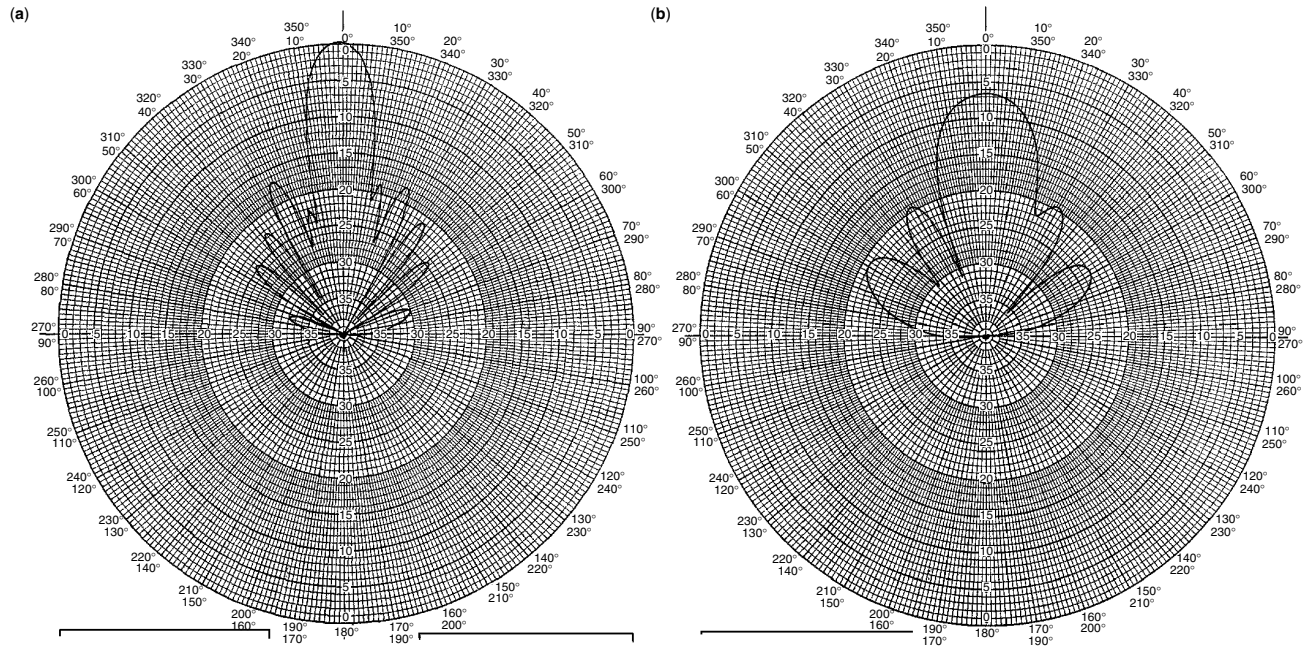


Figure 14. Radiation performance of 32-element combination array: (**a**) H plane; (**b**) E plane.

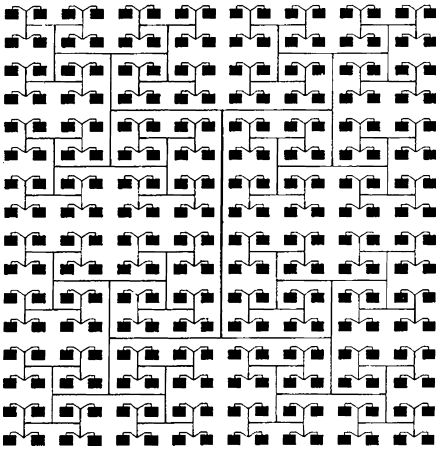


Figure 15. Schematic diagram of planar corporate fed array.

and microstrip technology can yield gains similar to those of a reflector for array sizes less than about 1000 elements, the feed-related losses (radiation, dielectric and ohmic) become significant.

There are printed alternatives to large arrays of patches to produce high-gain antennas for point-to-point applications. These include lens coupled printed antennas [e.g., 11] and reflectarrays [12]. Lens coupled microstrip patches remove the feed-associated losses as there is only one radiating element. These antennas can yield gains in excess of 30 dBi and importantly bandwidths (both radiation and impedance) as broad as the feed element [13]. Figure 16 shows a photograph of an aperture stacked patch lens coupled antenna, with a bandwidth that covers the entire Ka band (26–40 GHz). Printed reflectarrays are another promising alternative to large arrays of microstrip patches. These antennas can yield gains greater than 50 dBi, although the bandwidths are typically small, to date a couple of percent. Figure 17 shows a photograph of a millimeter wave reflectarray [14]. Of course, these printed antennas have many of the features of microstrip patch arrays; however, the conformal nature of the entire antenna is no longer a feature.

3.2. Scanning/Phased Planar Arrays

As mentioned previously, microstrip patch antennas can readily be integrated with active microwave devices. This is one of the key reasons as to why microstrip patch antennas are so advantageous when considering a scanning array, in particular a planar phased array. A planar phased array allows for pattern control in both dimensions and in doing so provides a very flexible or smart antenna.

Table 1. Planar Arrays of Microstrip Patches Versus Reflectors

Number of Elements	16	64	256	1024	4096
Directivity without network	20.9	27.0	33.0	39.2	45.1
Radiation loss	0.8	1.0	1.3	1.9	2.6
Surface wave loss	0.3	0.3	0.2	0.2	0.1
Dielectric loss	0.1	0.3	0.5	1.0	2.1
Ohmic loss	0.1	0.3	0.6	1.2	2.4
Calculated gain	19.5	25	30	34.5	37.5
Gain of reflector	18	24	30	36	42

Figure 16. Photograph of millimeter wave lens-coupled proximity-coupled microstrip patch antenna.

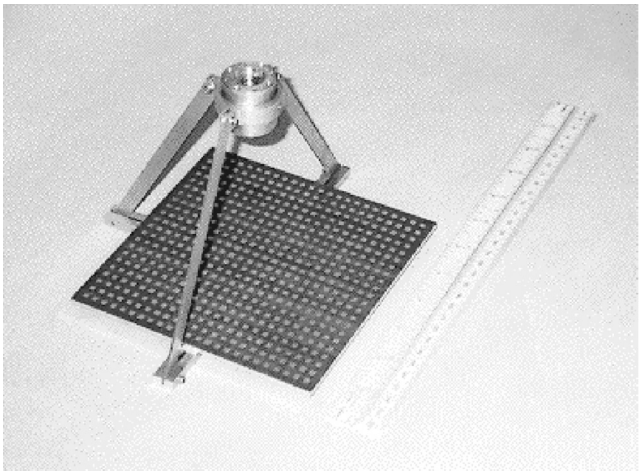


Figure 17. Photograph of millimeter wave reflectarray antenna.

There is one issue related to microstrip patch antenna technology that wasn't mentioned before in this article: surface wave excitation. Surface waves (sometime referred to as "leaky" waves) are "trapped" waves excited by the presence of the substrate or dielectric layers associated with the microstrip antenna. Because the energy is generally trapped within the material and not radiated,

surface waves are classified as a loss mechanism. The presence of a surface wave can cause increases in cross-polarization levels due to the trapped wave refracting off the finite edges of the ground plane of the antenna. Surface waves can also cause unwanted coupling between the antenna and any active devices.

For a single-layer microstrip patch antenna, the thicker the material used, the larger the power lost to the surface wave. Also the higher the dielectric constant, the less efficient the antenna becomes as a result of surface wave excitation. For large arrays of microstrip patches, the resonance of modes associated with these surface waves can severely limit the scan performance of the array by inducing a phenomenon known as a *scan blindness*. For a scan blindness, all (or at least most) of the power is coupled back to the source and subsequently is not radiated. A common means of examining the scanning potential of a large array of microstrip patches is to consider the theoretical active reflection coefficient of the array, which is defined as the reflection coefficient of an element in the array as a function of scan angle [15]. Figure 18 shows the active reflection coefficient of a large array of probe-fed patches when scanning in the E, H and D planes. As can be seen here, in the E plane, the active reflection becomes larger as the scan angle increases as a result of mutual coupling until a point where it levels off and then increases to unity at endfire. The scan angle where it is becomes large (approximately 75°) is the scan blindness. Note that the degree of blindness depends on what element is used. For example, if aperture-coupled patches were used in this array, the active reflection coefficient at the scan blindness would approach one. The magnitude at the scan blindness is dependent on the level of spurious radiation from the antenna.

The scan position of the blindness is very dependent on the element spacing of the array. Figure 19 shows the active reflection coefficient for an array of microstrip patches at three frequencies, at the lower edge of the 10-dB return loss bandwidth, the center frequency, and the upper frequency of the 10-dB return loss bandwidth. The element spacing of the array was 0.5 λ_0 at the center frequency. As can be seen from Fig. 19, the array has very limited scanning ability at the higher-frequency edge because of the impedance mismatch associated with the surface wave.

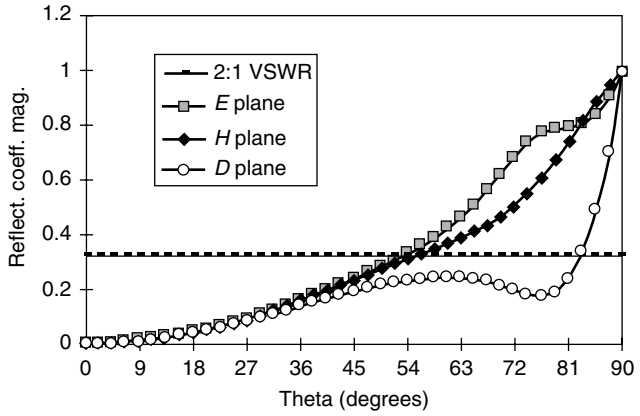


Figure 18. Scan active reflection coefficient of infinite array of probe-fed microstrip patches.

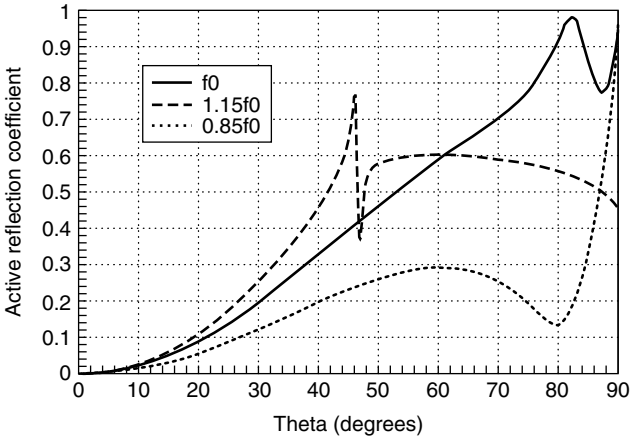


Figure 19. Frequency dependence of scan active reflection coefficient of infinite array of microstrip patches.

The sudden drop in reflection coefficient after the scan blindness is due to the presence of a grating lobe. Although the active reflection coefficient looks reasonable after this scan angle, the radiation efficiency of the array is low, due to power being dumped into the grating lobe [16,17]. Thus it would appear that microstrip patches would have very limited use for large scanning arrays because of the excitation of surface waves and the fact that to increase the bandwidth of a conventional patch the material thickness must be increased and therefore the surface wave content would also increase. However fortunately there are ways to alleviate this problem.

The previously mentioned scanning and/or material trends apply only to single-layer geometries and do not hold for more complicated patch configurations consisting of multiple layers. For example, an aperture stacked patch [18] can have a surface wave efficiency greater than 85% even when the overall thickness of the materials used is greater than $0.1\lambda_0$ (which is very thick for microstrip patches). Such a printed antenna can have an impedance bandwidth of over an octave. Also a stacked patch using high-dielectric-constant material for the lower layer can have an efficiency greater than 90%, even though a single-layer patch using the same high-dielectric-constant material has a surface wave efficiency of only <65% [19]. The 10-dB return loss bandwidth for this antenna can be greater than 30%. Both of these patch elements can also be used in large scanning arrays. It has since been was shown that a large array of aperture stacked patches can have a 10-dB return loss bandwidth in excess of an octave while being able to be scanned to angles greater than ±45° in the principal planes [20]. To design such arrays requires careful consideration of the impedance response of the array and its *spiders* (how the impedance changes as a function of scan angle [20]). It is imperative to try to minimize the impedance variation (as a function of frequency and scan angle) as much as possible for the array to ensure optimum performance.

Other techniques have been developed over the years that can improve the scanning performance of the conventional microstrip patch phased array, albeit at the expense of complexity. These include using cavity-backed

structures [21] and shorting pins [22]. These methods could be applied to the broadband solutions of Refs. 20 and 23 to give perhaps the ultimate microstrip patch phased arrays.

4. CONCLUSIONS

In this article, an overview of microstrip patch array technology has been presented. Various forms of linear arrays were discussed. Case studies were given and a comparison of the advantages and issues associated with each type of array were presented. Corporate fed arrays are probably the most versatile with the largest bandwidth, although these arrays suffer from higher feed loss than do series and combination arrays. Combination arrays can provide a relatively simple design procedure and also good radiation and bandwidth results. A linear phased array was also presented, and its scanning performance is summarized.

Planar fixed-beam and scanned arrays utilizing microstrip patches were also investigated. Once again, corporate feeding is probably the easiest means of forming a planar array, especially if scanning the beam is required. Surface waves associated with the dielectric materials can have detrimental effects on the scanning performance of a large array of patches, although several methods have been established to overcome this inherent problem. Also, the scan/materials trends for single-layer geometries do not necessarily hold for more complicated, broader bandwidth printed structures, which is very fortuitous. Finally, high-gain printed antenna alternatives to large arrays was briefly examined. These antennas are very suited to point-to-point applications.

From the arrays and trends presented, it should be apparent that microstrip patches will continue to be one of the preferred options when choosing an antenna for a communication system.

Acknowledgments

The authors would like to thank the following people for the valuable discussions and input into the design and realization of some of the arrays presented in this article: Dr. D. Chui, Dr. A. Hoorfar, Dr. A. Nirmalathas, Dr. D. Novak, Mr. W. Rowe, Dr. S. Targonski, and Mr. D. Welch.

BIOGRAPHIES

Rod Waterhouse (S'90–M'94–SM'01) received the degrees of BE (Hons), MEngSc (Research) and Ph.D. from the University of Queensland, Australia, in 1987, 1990, and 1994, respectively. In 1994 he joined the School of Electrical and Computer Engineering at the RMIT University, Melbourne, Australia. From mid–2000 to the beginning of 2001 he was a visiting professor at the Department of Electrical and Computer Engineering UCLA, California, for three months and then a visiting researcher in the Photonics Technology Branch at the Naval Research Laboratories, Washington D.C., for a another 3 months while on his sabbatical. In June 2001, he took a leave of absence from RMIT and joined Dorsal Networks, Columbia, Maryland. His research interests include printed antennas, optically distributed wireless systems, photonic devices and optical systems. He has published over 140 papers and has three patents in these areas. Dr. Waterhouse chaired the IEEE Victorian MTTS/APS Chapter from 1998–2001.

Kamran Ghorbani was born in Mashad, Iran, in 1966. He received his B.E. degree in communication and electronic engineering (first honor) from RMIT University, Melbourne, Australia, in 1995. He has completed his Ph.D. degree at RMIT in 2001. After working as a RF designer for AWA Defense Industries Adelaide, South Australia, he joined the RF and photonic research group at RMIT University in 1996. From 1999 to 2001 he worked as senior RF designer for a telecommunication company. He rejoined RF Photonic Group at RMIT in 2001, where he is currently a research fellow. His research interests include integrated optics, phased array antenna, and microwave system design.

BIBLIOGRAPHY

1. C. A. Balanis, *Antenna Theory: Analysis and Design*, 2nd ed., Wiley, New York, 1996.
2. J. R. James, P. S. Hall, and C. Wood, *Microstrip Antenna Theory and Design*, Peter Peregrinus, London, 1981.
3. D. M. Pozar and D. H. Schaubert, Comparison of three series fed microstrip array geometries, *IEEE Antennas Propagation Symp.*, Ann Arbor, MI, July 1993, pp. 728–731.
4. D. M. Pozar and B. Kaufman, Design considerations for low sidelobe microstrip arrays, *IEEE Trans. Antennas Propag.* **38**: 1176–1185 (Aug. 1990).
5. S. D. Targonski and D. M. Pozar, Design of wideband circularly polarized aperture coupled microstrip antennas, *IEEE Trans. Antennas Propag.* **41**: 214–220 (Feb. 1993).
6. A. Nirmalathas, C. Lim, D. Novak, and R. B. Waterhouse, Progress in millimeter-wave fiber-radio access networks (invited), *Ann. Telecommun.* **56**: 27–38 (Jan./Feb. 2001).
7. D. M. Pozar, The active element pattern, *IEEE Trans. Antennas Propag.* **29**: 1176–1178 (Aug. 1994).
8. J. Huang, A parallel-series-fed microstrip array with high efficiency and low cross-polarization, *Microwave Opt. Technol. Lett.* **5**: 230–233 (May 1992).
9. R. J. Mailloux, J. F. McIlvenna, and N. P. Kernweis, Microstrip array technology, *IEEE Trans. Antennas Propag.* **29**: 25–37 (Jan. 1981).
10. E. Levine, G. Malamud, S. Shtrikman and D. Treves, A study of microstrip array antennas with the feed network, *IEEE Trans. Antennas Propag.* **37**: 426–434 (April 1989).
11. L. Mall and R. B. Waterhouse, Millimeter-wave proximity-coupled microstrip antenna on an extended hemispherical dielectric lens, *IEEE Trans. Antennas Propag.* (in press).
12. D. M. Pozar, S. D. Targonski, and H. D. Syrigos, Design of millimeter-wave microstrip reflectarrays, *IEEE Trans. Antennas Propag.* **45**: 287–296 (Feb. 1997).
13. R. B. Waterhouse, D. Novak, A. Nirmalathas, and C. Lim, Broadband printed antennas for point-to-point and point-to-multipoint wireless millimetre-wave applications, *IEEE Antennas Propagation Symp.* Utah (USA), July 2000, pp. 1390–1393.

14. S. D. Targonski and R. B. Waterhouse, Microstrip reflectarray analysis and design techniques, *5th Australian Symp. Antennas*, Sydney, Australia, Feb. 1996, p. 20.

15. D. M. Pozar and D. H. Schaubert, Scan blindness in infinite arrays of printed dipoles, *IEEE Trans. Antennas Propag.* **32**: 602–610 (June 1984).

16. D. M. Pozar, Scanning characteristics of infinite arrays of printed antenna subarrays, *IEEE Trans. Antennas Propag.* **40**: 666–674 (June 1992).

17. D. Novak and R. B. Waterhouse, Impedance behaviour and scan performance of microstrip patch arrays configurations suitable for optical beamforming networks, *IEEE Trans. Antennas Propag.* **42**: 432–435 (March 1994).

18. S. D. Targonski, R. B. Waterhouse, and D. M. Pozar, Design of wideband aperture-stacked patch microstrip antennas, *IEEE Trans. Antennas Propag.* **46**: 1246–1251 (Sept. 1998).

19. R. B. Waterhouse, Stacked patches using high and low dielectric constant material combination, *IEEE Trans. Antennas Propag.* **47**: 1767–1771 (Dec. 1999).

20. R. B. Waterhouse, Design and performance of large arrays of aperture stacked patches, *IEEE Trans. Antennas Propag.* **49**: 292–297 (Feb. 2001).

21. F. Zavosh and J. T. Aberle, Infinite phased arrays of cavity-backed patches, *IEEE Trans. Antennas Propag.* **42**: 390–398 (March 1994).

22. R. B. Waterhouse, The use of shorting posts to improve the scanning range of probe-fed microstrip patch phased arrays, *IEEE Trans. Antennas Propag.* **44**: 302–309 (March 1996).

23. R. B. Waterhouse, Design and scan performance of large, probe-fed stacked microstrip patch arrays, *IEEE Trans. Antennas Propag.* (in press).

MICROWAVE WAVEGUIDES

M. HAJIAN
L. P. LIGTHART
Delft University of Technology
Delft, The Netherlands

1. INTRODUCTION

A waveguide is used to guide electromagnetic waves at microwave frequency regions, and an open-ended waveguide to radiate them; which one is used depends on the application. A waveguide usually consists of a hollow metal pipe whose cross section is rectangular or circular. In most applications the open-ended waveguide is used as a feed element for a large reflector antenna or as an antenna element in active or passive phased-array antennas. Waveguide antennas provide optimum RF performance, that is, high aperture efficiency, high polarization purity, and low VSWR. Apart from inherent wide-bandwidth characteristics, they have the unique feature of a highpass filter behavior. In general the waveguide antenna protects the receiver system from unwanted electromagnetic interference (EMI) at frequencies lower than the cutoff frequencies of the waveguide modes. The reasons for its popularity are:

- Robustness
- Wide bandwidth
- High polarization purity
- Easy to mount on surfaces of spacecraft or aircraft for space applications
- Standard for calibrating other antennas
- Element for protecting the fields of larger transmit antenna
- Easy to manufacture

Energy transport in a waveguide can be achieved through propagation of so-called electromagnetic wave modes. These modes are solutions of the Maxwell equations and satisfy the boundary conditions. Such a hollow waveguide has characteristic cutoff frequencies connected with the propagation modes. The propagation of these modes depends on the operational frequency. If the frequency of the signal entering the waveguide is higher than the cutoff frequency of a given mode, then the electromagnetic mode energy can be transported through the waveguide with minimum attenuation because of the conduction losses in the waveguide walls. If the frequency of the incoming signal is lower than the cutoff frequency of a given mode, then the electromagnetic mode field is attenuated to a very low value within a short distance. It is convenient to design the waveguide such that the electromagnetic energy can be guided through the mode with only the lowest cutoff frequency. This mode is called the *fundamental* or *dominant mode*.

2. WAVEGUIDE APPLICATIONS

The waveguide is a low-loss transmission line that can handle high power signals. Since losses increase with frequency, waveguide applications can be found in the microwave and millimeter-wave region. In telecommunication and radar systems where losses may give major problems waveguide components are attractive. In other applications such as those in satellite communications, ground stations, and radar, where high power is a necessary requirement, waveguide solutions become attractive because they satisfy high power-handling capabilities.

The shape of the waveguide and its inner structure can be reconfigured in order to realize passive microwave components such as filters, couplers, phase shifters or active components such as oscillators and amplifiers [1]. The design and measurement results of a multichannel waveguide power divider based on *H*-plane or *E*-plane geometry have been discussed [2]. A corrugated waveguide has been used [3] to realize a phase shifter as part of a high-power dual reflector array antenna. Yoneyama developed a transmit/receive system at 35 GHz based on the NRDW (nonradiative dielectric waveguide) [4]. In NRDW the millimeter-wave field is concentrated inside the dielectric and it propagates similar as in a metal waveguide, meaning that minimum "leakage" takes place.

The waveguide is extensively used in medical applications (e.g., cancer therapy) where electromagnetic energy is coupled into the human body [5]. For this purpose the waveguide is loaded with a lossless dielectric material

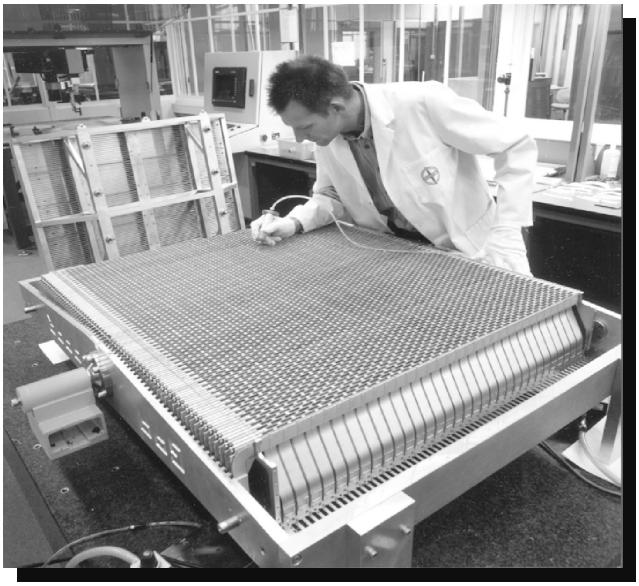


Figure 1. APAR open-ended waveguide array antenna under construction. (Courtesy of THALES.)

with a permittivity equal to that of the muscle tissue. This approach provides good impedance matching and results into a concentrated energy transfer.

The waveguide is often applied as an antenna element in large or phased-array antennas. For example, the active phased array antenna (APAR) from THALES uses open-ended waveguides as antenna elements [6,7]. APAR has four antenna-array panels; each panel consists of more than 4096 waveguide radiators (Fig. 1). Each waveguide radiator is connected to a T/R element, which comprises a sum channel and a combined transmit/delta elevation channel for monopulse tracking radar. The antennas are designed to have a wide angular scan range (up to 70° from the antenna broadside) for full 360° coverage, fast electronic beam steering to support search functions, and simultaneous tracking of hundreds of targets [8]. The APAR waveguide array uses a dielectric sheet for wide-angle impedance matching (WAIM sheet). Figure 2 shows an artist's impression of an antenna array panel integrated with the T/R modules and combiner networks.

It is possible to realize a linear array by using so-called slotted waveguides: narrow openings in the waveguide surface. A proper design of the slotted waveguide array may result in antennas with high efficiency, ultralow sidelobes and can sustain high peak power in the order of kilowatts. Figure 3 shows an example of a planar slotted waveguide array in X band.

The Earth Observation Satellites (ERS-1 and ERS-2) use a slotted waveguide array. The satellites were launched by the European Space Agency in 1991 and 1995, respectively. The ERS-1 antenna system under test can be seen in Fig. 4.

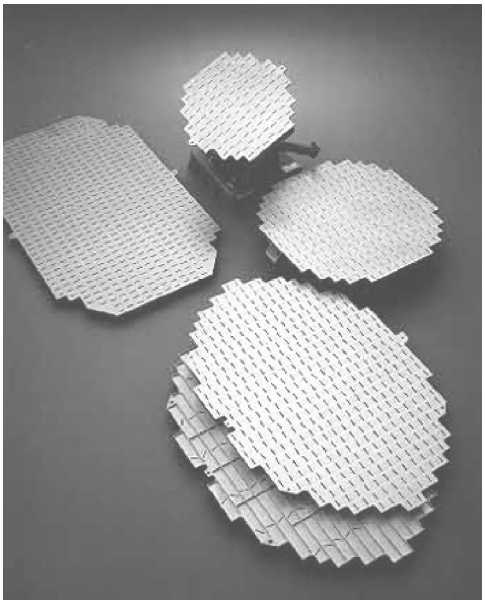


Figure 3. Slotted waveguide array. (Courtesy of ELTA Electronics Industries.)

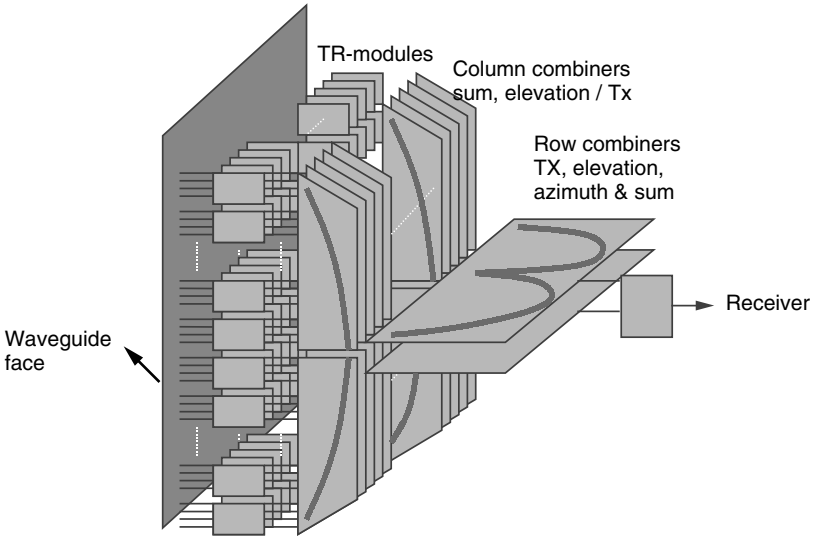


Figure 2. Artist impression of antenna RF network of APAR. (Courtesy of THALES.)

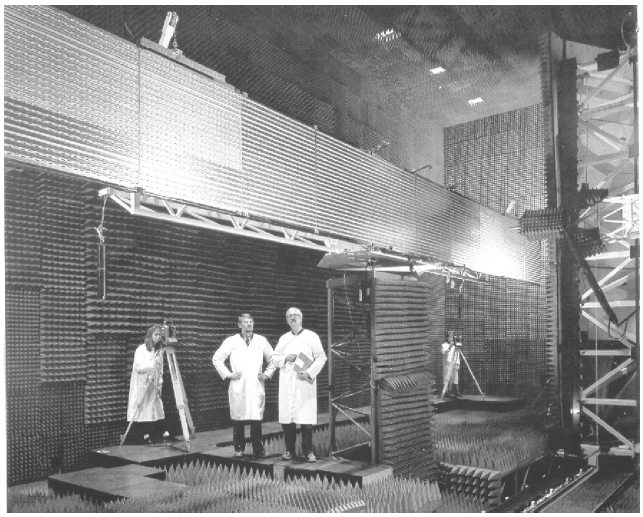


Figure 4. Slotted waveguide array antenna of the European remote-sensing satellite ERS-1 during planar near-field measurements. (Courtesy of Ericsson-ESA.)

Several applications of waveguide antennas in earth stations for satellite communication can be found in the literature. Bird et al. designed and measured [9] a compact high-power S-band dual frequency and dual polarized waveguide feed system with high power capability (handling 2 kW continuous RF) [9]. A second example is given by Bird and Sprey [10], who designed and measured a circularly polarized X-band feed system with high transmit/receive isolation for a dual-shaped Cassegrain reflector. A dual-mode waveguide filter and a 2-step waveguide *E*-plane filter has also been designed and measured [11,12].

The opening of the waveguide with different cross sections is tapered (flared) to a larger opening to form a so-called horn antenna. Such antennas are widely used as feed elements for large-sized radioastronomy, satellite and communication dishes. In the following an overview of different configurations with specific examples is given.

2.1. Single-Feed Systems

Figure 5 shows a selection of waveguide horn antennas for space applications.

2.1.1. TV-SAT Horn Antennas. The elliptical corrugated horn radiator is used in the TV-SAT feeding system in a reflector. The pattern has a high-gain elliptical beam (3-dB beamwidths $0.72^\circ \times 1.62^\circ$). The operational frequency is 11.7–12.1 GHz. It is circularly polarized and has low cross polarization (decoupling >35 dB).

2.1.2. IntelSAT 8 and Nahuel Horn Antennas. These conical corrugated horn antennas are part of the feed system of a dual-reflector Gregorian and a shaped reflector antenna in the frequency range of 10.95–14.5 GHz. They are linearly polarized and combine the transmit/receive function for both polarizations with low cross-polar coupling (decoupling >40 dB). The antennas can handle 12 high-power carriers up to 120 W per polarization.

2.2. Multifeed Systems

Figure 6 shows some multifeed antenna systems.

2.2.1. DFH-3 Feed. The seven-element diagonal-waveguide horn-antenna cluster combines a transmit (4 GHz) and receive (6 GHz) diplexer with a three-layer coaxial beamforming network (BFN).

2.2.2. AMOS 8 Feed. This three-element conical corrugated horn antenna cluster is part of an offset reflector antenna in the frequency range of 10.95–14.5 GHz.

2.2.3. IntelSAT 8 Feed Array. This 96-element conical corrugated horn antenna cluster as a feed system of a multibeam antenna provides eight beams with stringent

Figure 5. Waveguide horn antennas for space applications. (Courtesy of Astrium.)

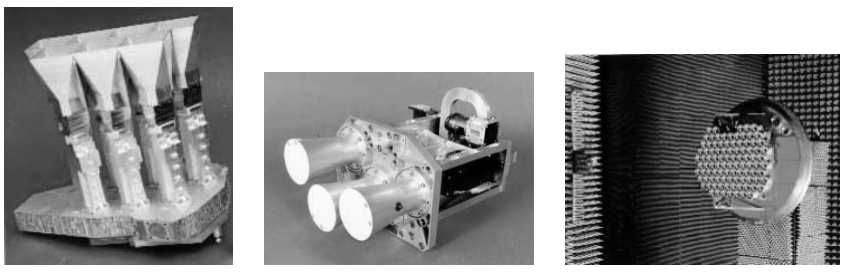


Figure 6. Multifeed waveguide systems for space applications. (Courtesy of Astrium.)

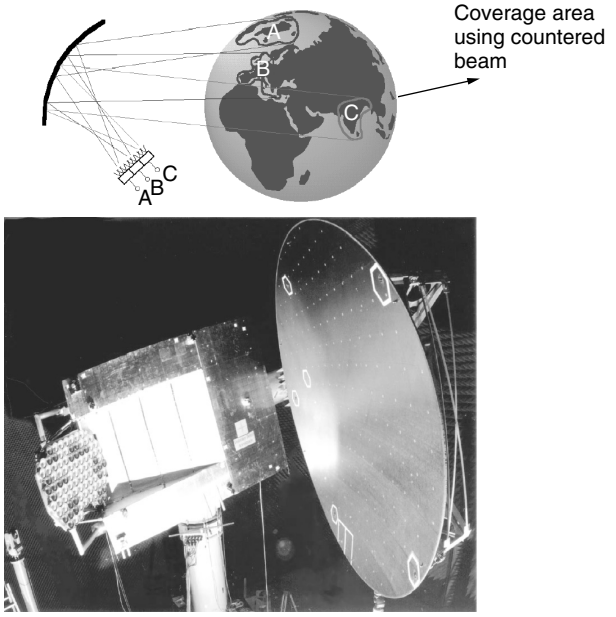


Figure 7. Reflector antenna with multiple contoured beams using the combined multifeed concept applied in Intelsat-8 (C band, transmitter and receiver) and in Intelsat-9 (C band, transmitter and receiver). (Courtesy of Astrium.)

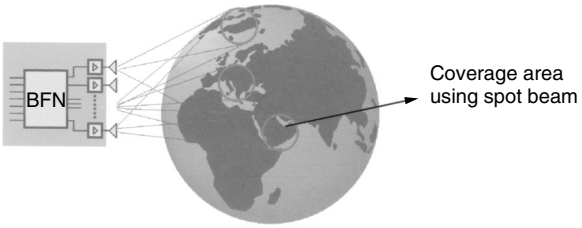
interbeam isolations (better than 27 dB) in order to support multifold frequency reuse. A three-layer coaxial beamforming network (BFN) generates eight individual beams: two left-hand circularly polarized hemispherical beams and six right-hand polarized "zone" beams. The power capability is 1.5 kW RF. Figure 7 illustrates the concept of generating multiple contoured beams. The same figure also shows the complete antenna system during the measurement phase.

Figure 8 shows a compact 8×8 dual-polarized waveguide array for application in a direct radiating antenna array. It was developed for space-based high-resolution polarimetric synthetic aperture radar (SAR) antennas in the X band. The height of the element including the dual polarized feed section is less than 0.3λ. A balun-type feed is used to excite the orthogonal fundamental modes with a polarization purity better than 40 dB over a bandwidth greater than 5%. This technique allows tight packaging in the array configuration and supports low-loss distribution/combiner networks.

In near-field antenna measurement techniques the waveguide antenna is used as a probe antenna to measure the radiation characteristics of the antenna under test (AUT). Since its characteristics are well measured and documented, correcting the probe to determine the far field accurately is more straightforward.

3. RECTANGULAR WAVEGUIDE

Figure 9 shows the cross section of a rectangular waveguide with a width and height of a and b respectively. It is assumed that the waveguide is filled with air and is infinite in length. There are a number of transverse electric- and magnetic modes (TE^x, TM^x, TE^y, TM^y, TE^z, TM^z) that satisfy the boundary conditions and are a solution to the Maxwell equations. The desired mode can be generated by the feed structure in the waveguide. This will be explained later in this article. However, without loss of generality in this part only TE^z is considered. Note that since the TEM mode does not satisfy the boundary conditions in the waveguide, it cannot be used to transport electromagnetic energy through this mode in the waveguide [11].



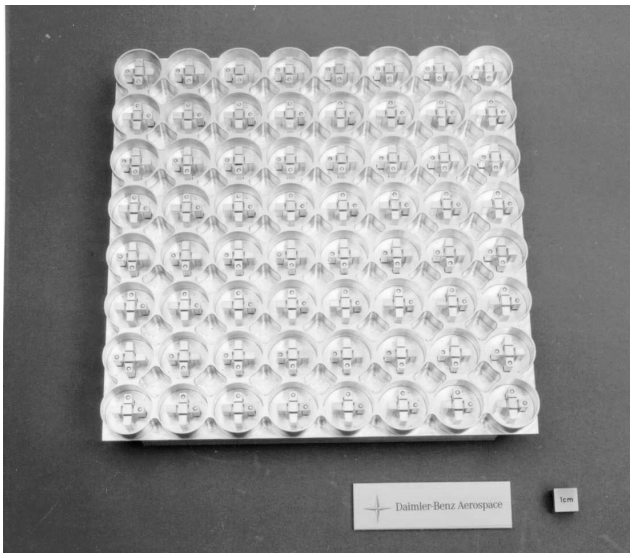


Figure 8. Open-waveguide planar array for space-based synthetic aperture radar (SAR) applications. (Courtesy of Astrium.)

3.1. Transverse Electric (TE^z)

Transverse electric modes are field configurations whose electric-field components lie in a plane that is transverse to the direction of the wave propagation. For example TE^z implies that $E_z = 0$. The other field components may or may not exist.

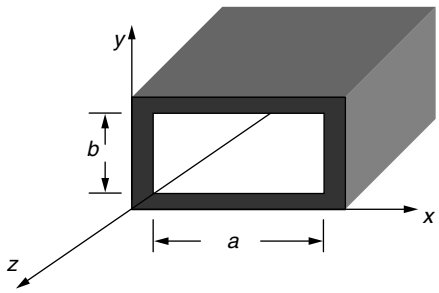


Figure 9. The rectangular waveguide with its dimensions and coordinates.

To derive the field expressions in a rectangular coordinate system that are TE to given direction, one needs only to let the magnetic vector potential $\mathbf{F}$ have only one component in that direction. Other components of electric vector potential $\mathbf{A}$, and $\mathbf{F}$ are set equal to zero. This corresponds to the following condition

$$\begin{aligned} \mathbf{A} &= 0 \\ \mathbf{F} &= \hat{a}_z F_z(x, y, z) \end{aligned} \tag{1}$$

where F_z is the scalar potential function and it represent the z component of vector potential $\mathbf{F}$. In the source-free region for the transverse electric modes, the components of the electric and magnetic fields satisfy the following equations [11]

$$\begin{aligned} E_x &= -\frac{1}{\varepsilon}\frac{\partial F_z}{\partial y} & H_x &= -j\frac{1}{\omega\mu\varepsilon}\frac{\partial^2 F_z}{\partial x\,\partial z} \\ E_y &= \frac{1}{\varepsilon}\frac{\partial F_z}{\partial x} & H_y &= -j\frac{1}{\omega\mu\varepsilon}\frac{\partial^2 F_z}{\partial y\,\partial z} \\ E_z &= 0 & H_z &= -j\frac{1}{\omega\mu\varepsilon}\left(\frac{\partial^2}{\partial z^2} + \beta^2\right)F_z \end{aligned} \tag{2}$$

where ε and μ are the permittivity in F/m and the permeability in H/m of the medium, respectively; ω is the frequency of the impressed signal; and β is the free-space wavenumber. The scalar potential F_z satisfies the scalar wave or Helmholtz equation

$$\nabla^2 F_z + \beta^2 F_z = 0 \tag{3}$$

In rectangular coordinates, this equation becomes

$$\frac{\partial^2 F_z}{\partial x^2} + \frac{\partial^2 F_z}{\partial y^2} + \frac{\partial^2 F_z}{\partial z^2} + \beta^2 F_z = 0 \tag{4}$$

The solution to Eqs. (3) and (4) is a well-known problem in the literature. The solution is based on the separation of variables and has the form of

$$F_z = \psi(x)\varphi(y)\zeta(z) \tag{5}$$

The variation in the z direction represents the propagating waves such that $\zeta(z)$ has the following form

$$\zeta(z) = A_1 e^{-j\beta_z z} + B_1 e^{+j\beta_z z} \tag{6}$$

where $\pm$ represents the waves traveling in the $+$ and $-z$ direction, respectively. It is assumed that the source in the waveguide is located such that only the waves in the $+z$ direction exist. In this case B_1 is zero.

The variation in the x and y directions represent the standing waves since the guide is bounded in these directions. The most appropriate solution is

$$\begin{aligned} \psi(x) &= A_2\cos(\beta_x x) + B_2\sin(\beta_x x) \\ \varphi(y) &= A_3\cos(\beta_y y) + B_3\sin(\beta_y y) \end{aligned} \tag{7}$$

where A_1, A_2, B_2, A_3, B_3 and $\beta_x, \beta_y, \beta_z$ are constants that need to be evaluated using the boundary conditions. $\beta_x, \beta_y, \beta_z$ are the wavenumbers in the x, y, and z directions, respectively, and they are related to the free-space wavenumber β in rad/m as follows:

$$\beta_x^2 + \beta_y^2 + \beta_z^2 = \beta^2 = \omega^2\mu\varepsilon \tag{8}$$

Substituting (6) and (7) in (5) with $B_1 = 0$ leads to

$$\begin{aligned} F_z(x, y, z) = {} & [A_2\cos(\beta_x x) + B_2\sin(\beta_x x)] \\ & * [A_3\cos(\beta_y y) + B_3\sin(\beta_y y)] * A_1 e^{-j\beta_z z} \end{aligned} \tag{9}$$

Equation (9) applies for $+z$ traveling waves. Note that here for simplicity the sign $+$ is omitted. Since the waveguide walls are good conductors, the tangential components of the electric field will vanish on the waveguide walls. For Fig. 9, the following boundary conditions exist for the left and right sidewalls:

$$\begin{aligned} E_y(x = 0, 0 \le y \le b, z) &= E_y(x = a, 0 \le y \le b, z) = 0 \\ E_z(x = 0, 0 \le y \le b, z) &= E_z(x = a, 0 \le y \le b, z) = 0 \end{aligned} \tag{10}$$

and for the top and bottom walls

$$\begin{aligned} E_x(0 \le x \le a, y = 0, z) &= E_x(0 \le x \le a, y = b, z) = 0 \\ E_z(0 \le x \le a, y = 0, z) &= E_z(0 \le x \le a, y = b, z) = 0 \end{aligned} \tag{11}$$

Equations (2), (10), and (11) are used to determine the constants in Eq. (9). Substituting (9) in (2), the y component of the electric field can be written as

$$\begin{aligned} E_y(x, y, z) = {} & \frac{\beta_x}{\varepsilon}[-A_2\sin(\beta_x x) + B_2\cos(\beta_x x)] \\ & * [A_3\cos(\beta_y y) + B_3\sin(\beta_y y)] * A_1 e^{-j\beta_z z} \end{aligned} \tag{12}$$

Applying the boundary condition given by Eq. (11) on the left wall for the E_y component to Eq. (12) gives

$$\begin{aligned} E_y(x = 0, 0 \le y \le b, z) = {} & \frac{\beta_x}{\varepsilon}[B_2] \\ & * [A_3\cos(\beta_y y) + B_3\sin(\beta_y y)] \\ & * A_1 e^{-j\beta_z z} = 0 \end{aligned} \tag{13}$$

Equation (13) can be satisfied if and only if B_2 is equal to zero. Applying the boundary condition of the right wall to Eq. (12) leads to

$$\begin{aligned} E_y(x = a, 0 \le y \le b, z) = {} & \frac{\beta_x}{\varepsilon}[-A_2\sin(\beta_x a)] \\ & * [A_3\cos(\beta_y y) + B_3\sin(\beta_y y)] \\ & * A_1 e^{-j\beta_z z} = 0 \end{aligned} \tag{14}$$

Equation (14) can be satisfied for a nontrivial solution if and only if

$$\sin(\beta_x a) = 0, \quad \beta_x a = m\pi \quad m = 0, 1, 2, \ldots \tag{15a}$$

$$\beta_x = \frac{m\pi}{a} \quad m = 0, 1, 2, \ldots \tag{15b}$$

Usually Eqs. (15a) and (15b) are called the *eigenfunction* and *eigenvalue*. It is straightforward to show that the following relations exists, using the same procedure

$$\begin{aligned} B_3 &= 0 \\ \beta_y &= \frac{n\pi}{b} \quad n = 0, 1, 2, \ldots \end{aligned} \tag{16}$$

Substituting Eqs. (16) and (15) in (9) and letting $A_1A_2A_3 = A$ leads to

$$F_z(x, y, z) = A \cos\left(\frac{m\pi}{a}x\right) \cos\left(\frac{n\pi}{b}y\right) e^{-j\beta_z z} \tag{17}$$

Substituting (17) in (2) leads to the complete solution of TE^z_{mn} modes

$$\begin{aligned} E_x &= A\frac{\beta_y}{\varepsilon} \cos(\beta_x x) \sin(\beta_y y) e^{-j\beta_z z} \\ E_y &= A\frac{\beta_x}{\varepsilon} \sin(\beta_x x) \cos(\beta_y y) e^{-j\beta_z z} \\ E_z &= 0 \\ H_x &= A\frac{\beta_x\beta_z}{\omega\mu\varepsilon} \sin(\beta_x x) \cos(\beta_y y) e^{-j\beta_z z} \\ H_y &= A\frac{\beta_z\beta_y}{\omega\mu\varepsilon} \cos(\beta_x x) \sin(\beta_y y) e^{-j\beta_z z} \\ H_z &= -jA\frac{\beta_x^2 + \beta_y^2}{\omega\mu\varepsilon} \cos(\beta_x x) \cos(\beta_y y) e^{-j\beta_z z} \end{aligned} \tag{18}$$

where β_x and β_y are the wavenumbers (eigenvalues) in the x and y directions, respectively. They are related to the wavelengths of the wave inside the waveguide in the x and y directions and the wave number in the z direction β_z and β as follows:

$$\begin{aligned} \beta_y &= \frac{n\pi}{b} = \frac{2\pi}{\lambda_y} \quad n = 0, 1, 2, \ldots \\ \beta_x &= \frac{m\pi}{a} = \frac{2\pi}{\lambda_x} \quad m = 0, 1, 2, \ldots \\ \beta_z^2 &= \beta^2 - (\beta_x^2 + \beta_y^2) = \beta^2 - \left[\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2\right] \\ \beta &= \frac{2\pi}{\lambda} \end{aligned} \tag{19}$$

The values of β_z depend on the waveguide cutoff frequency and its value determines the propagating waves, standing waves, and evanescent waves. The cutoff frequency and cutoff wavenumber in turn are determined by letting $\beta_z = 0$ in Eq. (8) or (19)

$$\beta_c^2 = \beta^2 = \omega^2\mu\varepsilon = \omega_c^2\mu\varepsilon = (\beta_x^2 + \beta_y^2) = \left[\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2\right] \tag{20}$$

which leads to

$$\begin{aligned} 2\pi f_c\sqrt{\mu\varepsilon} &= \sqrt{\left[\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2\right]} \\ &\times \begin{cases} (f_c)_{mn} = \dfrac{1}{2\pi\sqrt{\mu\varepsilon}}\sqrt{\left[\left(\dfrac{m\pi}{a}\right)^2 + \left(\dfrac{n\pi}{b}\right)^2\right]} \\ \qquad \times \left\{\begin{matrix} n = 0, 1, 2, \ldots \\ m = 0, 1, 2, \ldots \end{matrix}\right\} m = n \neq 0 \end{cases} \end{aligned} \tag{21}$$

Since it is assumed that the z direction is the propagation axis, the integers m and n denote the number of half-waves of electric or magnetic field intensity in the x and y directions. Depending on the value of the cutoff frequency, the propagation constant β_z can take different values. Three different cases are distinguished and they are given here as follows:

$$\beta_c^2 \triangleq \beta_x^2 + \beta_y^2 = \beta^2 - \beta_z^2$$

where

$$\beta_z = \begin{cases} \pm\sqrt{\beta^2 - \beta_c^2} = \pm\beta\sqrt{1 - \left(\dfrac{f_c}{f}\right)^2} \\ \qquad \text{for } f > f_c \text{ propagating waves} \\ 0 \qquad \text{for } f = f_c \text{ standing waves} \\ \pm j\sqrt{\beta_c^2 - \beta^2} = \pm j\beta\sqrt{\left(\dfrac{f_c}{f}\right)^2 - 1} \\ \qquad \text{for } f < f_c \text{ evanescent waves} \end{cases} \tag{22}$$

The evanescent waves are the fields that decay exponentially. Equation (22) also shows that the waveguide has highpass filter behavior. If the operational frequency is higher than the cutoff frequency, the fields propagate; if not, they attenuate. The ratio of suitable electric to magnetic field components has the same dimension as the impedance. Using (18) the following relation exists:

$$Z \triangleq \frac{E_x}{H_y} = -\frac{E_y}{H_x} = \frac{\omega\mu}{\beta_z}\Omega \tag{23}$$

Inserting Eq. (22) in (23) leads to the following expression for the waveguide impedance

$$Z = \begin{cases} \dfrac{\eta}{\sqrt{1 - \left(\dfrac{f_c}{f}\right)^2}} & \text{for } f > f_c \text{ resistive} \\ \infty & \text{for } f = f_c \text{ open circuit} \\ j\dfrac{\eta}{\sqrt{\left(\dfrac{f_c}{f}\right)^2 - 1}} & \text{for } f < f_c \text{ inductive} \end{cases} \tag{24}$$

where η is the free-space impedance. The waveguide impedance behaves inductively for frequencies lower than the cutoff frequency, and resistively for frequencies higher than the cutoff frequency. Figure 10 shows the waveguide impedance as function of the normalized frequency.

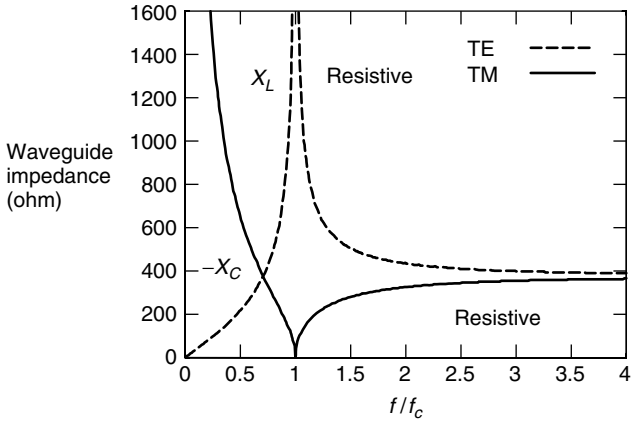


Figure 10. The wave impedance of a rectangular waveguide.

The expression for the wavenumber β_z along the z axis can be used to define the wavelength along the guide axis and is given as

$$\lambda_z = \begin{cases} \dfrac{\lambda}{\sqrt{1-\left(\dfrac{f_c}{f}\right)^2}} = \dfrac{\lambda}{\sqrt{1-\left(\dfrac{\lambda}{\lambda_c}\right)^2}} & \text{for } f > f_c \\ \infty & \text{for } f = f_c \\ j\dfrac{\lambda}{\sqrt{\left(\dfrac{f_c}{f}\right)^2-1}} = \dfrac{\lambda}{\sqrt{1-\left(\dfrac{\lambda}{\lambda_c}\right)^2}} & \text{for } f < f_c \end{cases} \tag{25}$$

Figure 11 shows the waveguide, wavelength, and wavenumber along the guide axis as function of the normalized frequency. Note that depending on the value of the signal and cutoff frequency, the waves are propagating or attenuating. The expression for the waveguide impedance for the TM mode is not given in this article. A procedure similar to TE can be used to derive the related parameters [11].

Example 1. The waveguide in Fig. 9 has inner dimensions of $a = 2$ cm, $b = 1$ cm, is filled with air and operates in the TE_{10} mode.

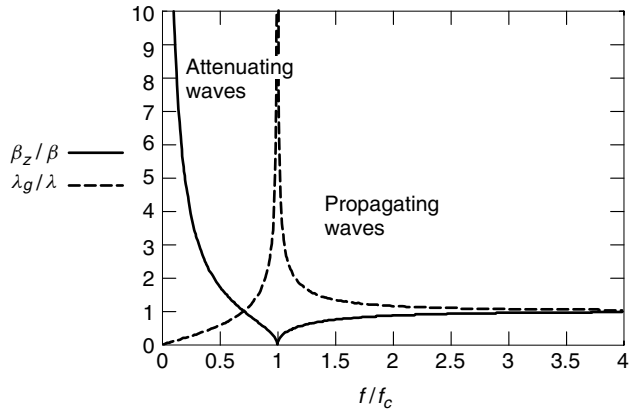


Figure 11. The normalized wavelength and propagation constant.

a. Determine the cutoff frequency.

b. Determine the propagation constant at 9.5 GHz.

c. Determine the waveguide impedance at 7.0 and 9.5 GHz.

Now the waveguide is filled with material. The dielectric constant of the material equals 5.

d. Find the cutoff frequency.

e. Determine the new waveguide dimensions to obtain the same cutoff frequency as derived in part a.

Solution

a. $(f_c)_{10} = \dfrac{c}{2a} = \dfrac{3\times10^8}{2\times2\times10^{-2}} = 7.5 \text{ GHz}$

b. $\beta_z = \beta\sqrt{1-\left(\dfrac{f_c}{f}\right)^2} = \dfrac{2\pi}{\lambda_0}\sqrt{1-\left(\dfrac{f_c}{f}\right)^2}$

$$= \frac{2\pi}{\dfrac{3\times10^8}{9.5\times10^9}}\sqrt{1-\left(\frac{7.5}{9.5}\right)^2} = 122.12 \text{ rad/m}$$

c. For 7.0 GHz:

$$Z = j\frac{\eta}{\sqrt{\left(\dfrac{f_c}{f}\right)^2-1}} = j\frac{120\pi}{\sqrt{\left(\dfrac{7.5}{7}\right)^2-1}}$$

$$= j980\ \Omega \text{ (inductive)}$$

For 9.5 GHz:

$$Z = \frac{\eta}{\sqrt{1-\left(\dfrac{f_c}{f}\right)^2}} = \frac{120\pi}{\sqrt{1-\left(\dfrac{7.5}{9.5}\right)^2}}$$

$$= 614\ \Omega \text{ (resistive)}$$

d. $(f_c)_{10} = \dfrac{c}{2a\sqrt{\varepsilon_r}} = \dfrac{3\times10^8}{2\times2\times10^{-2}\sqrt{5}} = 3354 \text{ GHz}$

e. $(f_c)_{10} = \dfrac{c}{2a\sqrt{\varepsilon_r}} = \dfrac{3\times10^8}{2\times a\sqrt{5}} = 7.5\times10^9, a = 8.94 \text{ mm}$

In the second part of Example 1 a miniaturization aspect of waveguides is introduced. In Section 4 the theory and practice of miniaturization and matching of a dielectric-filled waveguide will be discussed.

3.2. Power in Rectangular Waveguide

The power transport is associated with the fields propagating in the waveguide. The total power in the waveguide is the summation of the power of the TE_{mn} and TM_{mn} modes and is given as

$$P_{\text{total}} = \sum_m \sum_n P_{mn}^{\text{TE}} + \sum_m \sum_n P_{mn}^{\text{TM}} \tag{26}$$

In this section the expression for P_{mn}^{TE} is derived. A similar procedure can be used to derive P_{mn}^{TM}. The power is calculated by integrating the power density related to the

electromagnetic fields over the cross-sectional area of the waveguide and is given by

$$P_{mn} = \iint_{S_0} W_{mn} \cdot dS = \iint_{S_0} W_{mn} \cdot \hat{\boldsymbol{n}}\, ds$$
$$= \frac{1}{2} \iint_{S_0} \mathrm{Re}[\boldsymbol{E} \times \boldsymbol{H}^*]_{mn} \cdot \hat{\boldsymbol{n}}\, ds \tag{27}$$

where W_{mn} is the power density related to mn^{th} mode and $dS = dx\,dy$ is the infinitesimal area of the waveguide cross section and $\hat{\mathbf{n}} = \hat{n}_z$ is the unit vector normal to the waveguide cross section. The power density is given by

$$W_{mn} = \tfrac{1}{2}\mathrm{Re}[\boldsymbol{E} \times \boldsymbol{H}^*]_{mn} = \tfrac{1}{2}\mathrm{Re}[(\hat{n}_x E_x + \hat{n}_y E_y) \times (\hat{n}_x H_x + \hat{n}_y H_y)^*] = \tfrac{1}{2}\hat{n}_z \mathrm{Re}[E_x H_y^* - E_y H_x^*] \tag{28}$$

Inserting the given field components given by Eq. (18) in (28) and using (27) leads to the following expression for the power:

$$P_{mn} = \frac{1}{2}\int_0^a \int_0^b \hat{n}_z \mathrm{Re}[E_x H_y^* - E_y H_x^*]_{mn} \cdot \hat{n}_z\, dx\, dy$$
$$= \frac{1}{2}\int_0^a \int_0^b \mathrm{Re}[E_x H_y^* - E_y H_x^*]_{mn}\, dx\, dy$$
$$= \frac{1}{2}|A|^2 \frac{\beta_z}{\omega\mu\varepsilon^2} \int_0^a \int_0^b [\beta_y^2 \cos^2(\beta_x x)\sin^2(\beta_y y) + \beta_x^2 \cos^2(\beta_y y)\sin^2(\beta_x x)]\, dx\, dy \tag{29}$$

Performing the integration and inserting the expression for the β_z given by Eq. (22) in the propagating case leads to the final expression for the power transport by the TE mode as

$$P_{mn} = \frac{1}{2}|A|^2 \frac{\beta_z}{\omega\mu\varepsilon^2}(\beta_y^2 + \beta_x^2)\left(\frac{a}{\delta_m}\right)\left(\frac{b}{\delta_n}\right)$$
$$= \frac{1}{2}|A|^2 \frac{\beta(\beta_y^2 + \beta_x^2)}{\omega\mu\varepsilon^2}\left(\frac{a}{\delta_m}\right)\left(\frac{b}{\delta_n}\right)\sqrt{1 - \left(\frac{f_c}{f}\right)^2} \tag{30}$$

where

$$\delta_k = \begin{cases} 1 & k = 0 \\ 2 & k \neq 0 \end{cases} \tag{31}$$

Example 2. The waveguide in Fig. 9 has inner dimensions $a = 2$ cm, $b = 1$ cm, is filled with air and operates in the TE_{10} mode. The frequency is 9.5 GHz. The peak value of the electric field is 40 kV/m. Calculate the transport power in the waveguide.

Solution Since the waveguide operates in TE_{10} mode, the field components can be found using Eqs. (18) and (19) with $m = 1$ and $n = 0$. They are given by

$$E_x = 0$$
$$E_y = A\frac{\beta_x}{\varepsilon}\sin\left(\frac{\pi}{a}x\right)e^{-j\beta_z z}$$
$$E_z = 0$$
$$H_x = A\frac{\beta_x\beta_z}{\omega\mu\varepsilon}\sin\left(\frac{\pi}{a}x\right)e^{-j\beta_z z} \tag{32}$$
$$H_y = 0$$
$$H_z = -jA\frac{\beta_x^2 + \beta_y^2}{\omega\mu\varepsilon}\cos\left(\frac{\pi}{a}x\right)$$

The peak value of the electric field intensity can be obtained from the maximum electric field value by using equation (32) and is given by

$$|E_y|_{\max} = \frac{|A|}{\varepsilon_0}\beta_x = \frac{|E_{10y}|}{\varepsilon_0}\frac{\pi}{a} = 40 \text{ kV/m} \tag{33}$$

The cutoff frequency of the TE_{10} mode is

$$f_c = \frac{c}{2a} = \frac{3\times 10^8}{2\times 2\times 10^{-2}} = \frac{3\times 10^{10}}{4} = 7.5 \text{ GHz} \tag{34}$$

Inserting Eqs. (33) and (34) into Eq. (30) leads to the maximum power

$$P_{10} = \frac{|E_y|^2}{2\eta}\left(\frac{a}{\delta_{01}}\right)\left(\frac{b}{\delta_{00}}\right)\sqrt{1 - \left(\frac{f_c}{f}\right)^2}$$
$$P_{10} = \frac{|40\times 10^3|^2}{2\times 377}\frac{2\times 10^{-2}}{2}\frac{1\times 10^{-2}}{1}\sqrt{1 - \left(\frac{7.5}{9}\right)^2} \tag{35}$$
$$P_{10} \simeq 117.30 \text{ W}$$

3.3. Excitations of Modes in a Rectangular Waveguide

The analysis given in the previous sections focused on the wave propagation and power transport in the waveguide. However, the electric and magnetic fields first need to be generated in the waveguide. In general this can be done by an infinitesimal electric or magnetic dipole element (probe) [12], which in turn is connected to a generator. In order to achieve the optimal interface, it is necessary to match the impedance of the feed element to the waveguide impedance. This means that the return loss should be minimized. It is also desired that the dielectric losses, and losses caused by sharp bends are as low as possible.

The position, dimensions, and depth of the probe play a major role in coupling the energy from the feedline into the waveguide. The reflection caused by the waveguide walls and the generated field at the antenna probe need to be in phase in order to reinforce each other and to allow the waves to propagate in the aperture direction. In most cases the distance between the probe and the short-circuited end wall of the waveguide is in the order

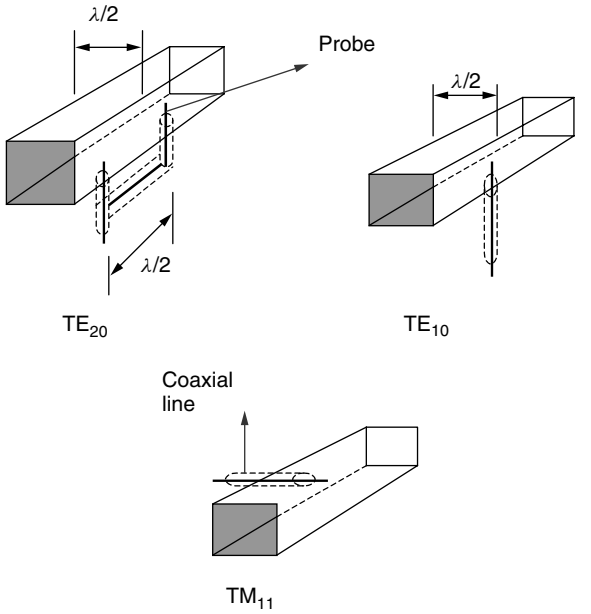


Figure 12. Methods used to excite various modes in a rectangular waveguide.

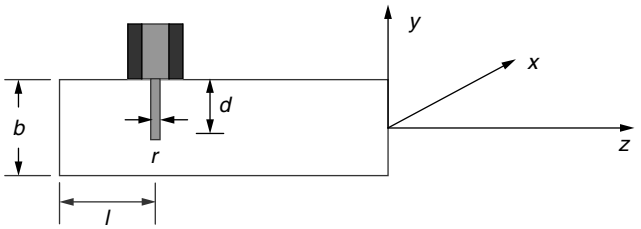


Figure 13. Configuration of coax-to-rectangular waveguide transition.

of half a wavelength. In this way the generated fields and reflected fields coming from the backside of the end wall are in phase. Figure 12 shows various methods to excite different modes in the waveguide. Figure 13 shows the geometry of a coaxial cable-to-waveguide transition. It consists of a coaxial line with its inner conductor extending over a distance d into the waveguide. To create a probe that radiates into one direction, a short is placed at a distance of l from the probe. By choosing l and d properly one can couple the power optimally from the coaxial line to the waveguide. r is the radius of the inner conductor of the probe.

The input impedance is inductive. Tian et al. [14] show that to obtain an optimal transition from coax to waveguide, one needs to introduce a capacitance between the coax end and the waveguide wall. When designing a coaxial feed, the major design problem is to find the optimal location and dimensions of the probe to achieve the best impedance matching. Equation (36) suggests that by properly choosing the probe length d and the short-circuit end position l, the radiation resistance R_{10} can be made equal to the characteristic impedance Z_0 of the coaxial line. In turn X can cancel the input reactance caused by the higher-order modes. The diameter of the coaxial feed is determined experimentally. The input impedance of the coaxial line is derived using the mode matching technique and is given by [13]

$$Z_{\text{in}} = R + jX$$

$$R_{10} = \frac{2Z_0}{ab\beta_{10}\beta}\sin^2(\beta_{10}l)\tan^2\left(\beta\frac{d}{2}\right)$$

$$X = \frac{Z_0}{ab\beta_{10}\beta}\tan^2\left(\beta\frac{d}{2}\right) \times \left\{\begin{array}{l} \ln\dfrac{2a}{\pi r} + \dfrac{0.0518\beta^2a^2}{\pi^2} \\ \quad + \dfrac{2\pi}{\beta_{10}a}\sin(2\beta_{10}l) \\ -2\left(1-\dfrac{2r}{a}\right) - 2\beta^2\displaystyle\sum_{m=1}^{\infty}\left[1-\frac{\sin^2\left(\dfrac{m\pi d}{2b}\right)}{\sin^2\left(\dfrac{\beta d}{2}\right)}\right] \\ \quad \times \dfrac{K_0(k_m r)}{k_m^2} \end{array}\right\}$$

$$k_m^2 = \left(\frac{m\pi}{b}\right)^2 - \beta^2$$

where a and b are the width and height of the waveguide, Z_0 is the free-space impedance, K_0 is the Bessel function of the second kind, and $\beta_{10} = \sqrt{\beta^2 - (\pi/a)^2}$ is the propagation constant of the fundamental mode. Figure 14 shows the values of l and d for tuning the input impedance of the probe for an X-band waveguide. The dimensions of the waveguide are a = 2.286 cm and b = 1.016 cm. The values make the inductive part of the input impedance equal to zero and force the resistive part to different characteristic impedances.

Figure 15 shows the optimum matching and measured input reflection of the coax transition of a dielectric filled waveguide in the L-band for $a = 83$ mm and $b = 10$ mm with a central frequency of 1.6 GHz. Since the height of the waveguide is very low, the waveguide behaves more

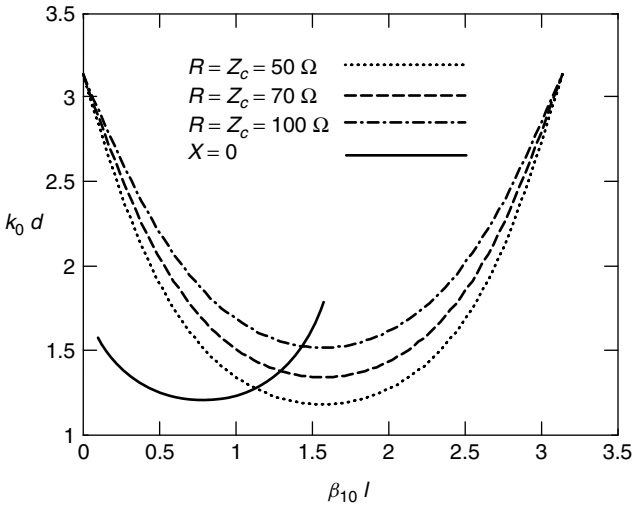


Figure 14. The design contour for matching the input impedance of the probe feed. (*Source*: R. E. Collin, *Field Theory of Guided Waves*.)

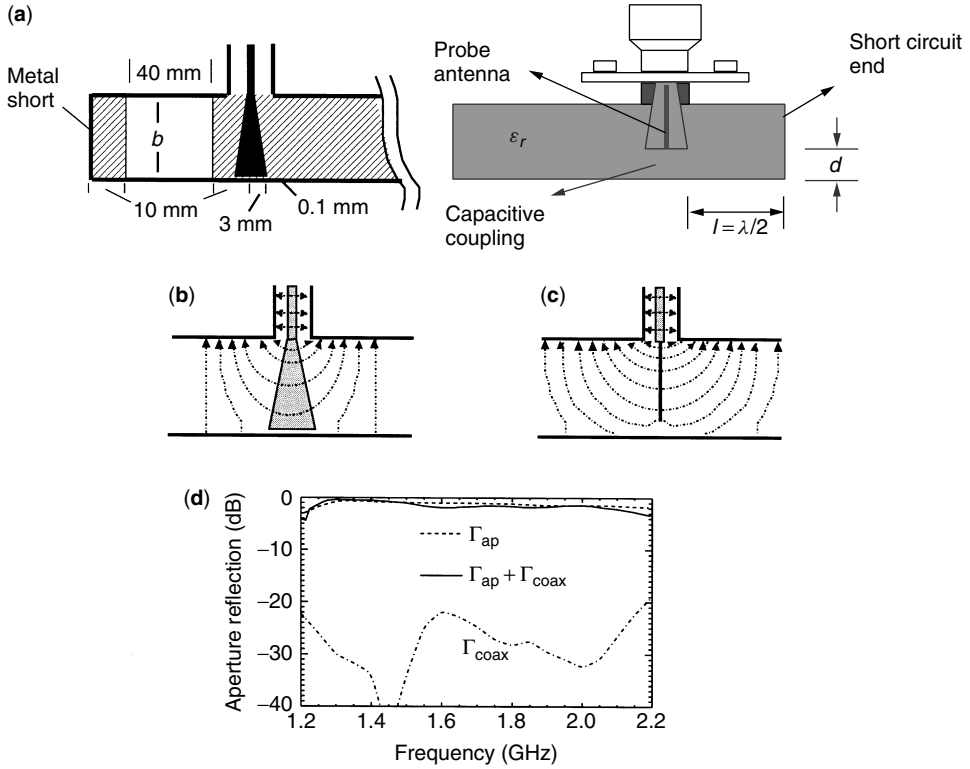


Figure 15. L-band coax-to-waveguide transition: (**a**) geometry of optimized capacitive coupling; (**b,c**) higher-order modes for a cone-shaped probe and monoprobe; (**d**) reflection matching of the coaxial waveguide transition, where Γ_{ap} and Γ_{coax} are the aperture and the input reflection at the coaxial interface.

or less as a cavity resonator. The probe feed needs to be tapered to a disk-cone form in order to increase the capacitive coupling and radiation resistance [14].

In many microwave and millimeter-wave planar circuit applications, such as active phased arrays or front ends in radar and radiocommunication systems, it is often necessary to use a microstrip line to excite the waveguide antenna [15]. Care is needed to couple the field generated at the source via the feed structure into the waveguide. The transition between the coaxial or microstrip feed is complex and needs to be analyzed, designed and experimentally verified.

There are several possible alternative transitions from microstrip to waveguide: microstrip *E*-plane probe (MEPP), finline transition, microstrip end launcher (MEL), ridged-waveguide transition, radiating-slot transition, and tapered-microstrip transition.

3.3.1. Microstrip *E*-Plane Probe (MEPP). Figure 16 shows the configuration of the MEPP. The probe behaves like a monopole antenna, which excites the waveguide. The reflector is used to reflect the excited waves in the desired direction.

A theoretical model has been developed [16] for calculating the input impedance. The model is based on an assumed current distribution in the probe, and a variational expression to calculate the input impedance.

Experimental results show that the MEPP has a −20 dB bandwidth of about 30% in the Ka band (26.5–40 GHz).

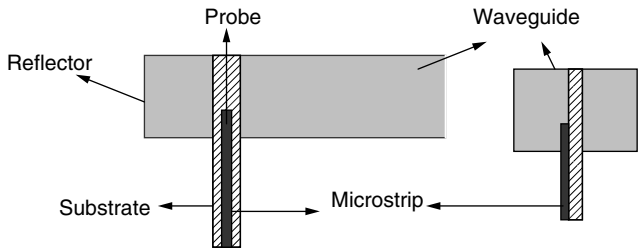


Figure 16. Microstrip *E*-plane probe (MEPP).

An advantage of MEPP is obviously its wide bandwidth; a disadvantage, however, is its non-planar construction. In the MEPP configuration, the microstrip T/R module would lie transversely to the waveguide, which is not very practical for miniature phased-array systems. Nevertheless, MEPP is one of the most widely used microstrip-to-waveguide transitions because of its simple configuration and its wide bandwidth.

3.3.2. Finline Transition. Figure 17 illustrates the finline transition where tapered antipodal fins are used to rotate the dominant TE_{10} mode of the waveguide into the TEM mode of the microstrip line. This transition does not require a reflector, since the bifurcation of the waveguide due to the microstrip ground plane serves as an imaginary reflector.

In Ref. 17 experimental results show a −20 dB bandwidth of 25% in the band 18–26 GHz. Unfortunately, this

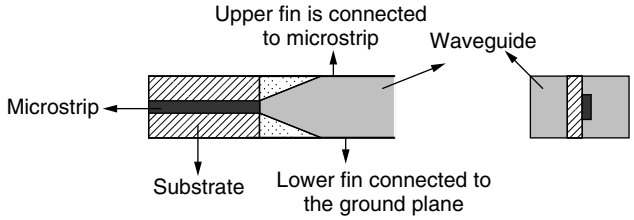


Figure 17. The geometry of finline transition.

paper does not present a theoretical model for analyzing such a transition. In Ref. 18 an empirical model based on a T matrix of a uniform finline section is given. The tapered finline is divided into a large number of uniform finlines. The total T matrix of the transition is calculated by taking the product of the T matrices of the individual uniform sections. In Ref. 19 an empirical expression for the resonance frequencies that may occur in this type of transition is given. By using this expression, one can place the resonance frequencies outside the operational frequency band in the design stage.

Advantages of this transition are the wide bandwidth and its configuration, which is suitable for miniature array systems. Disadvantages are its long complex structure and its empirical design.

3.3.3. Microstrip End Launcher (MEL). The MEL uses a loop antenna (launcher) to excite the waveguide (Fig. 18). A reflector is needed to reflect the excited waves into the desired direction.

In Ref. 20 a theoretical model has been derived to calculate the input impedance of the MEL. The model is based on an assumed current distribution in the launcher, and a variational expression for the input impedance. The experimental results of the MEL show a -20 dB bandwidth of 10% in the Ka band. The advantage of the MEL is its simple longitudinal configuration, which is suitable for miniature arrays. The disadvantage is that the model requires a rather narrow current strip (0.185 mm in the Ka band) of the launcher, and as a result only thin microstrip substrates can be used.

3.3.4. Ridged Waveguide Transition. Figure 19 shows the configuration of the ridged waveguide transition where a tapered or stepped ridge in a waveguide is used to convert the dominant TE_{10} mode of the waveguide into the TEM mode of the microstrip line.

In Ref. 21 the experimental result of such a transition was presented and it shows -20-dB difference over 25% bandwidth in the Ka band. Because of the complex structure of this transition, the final design was found

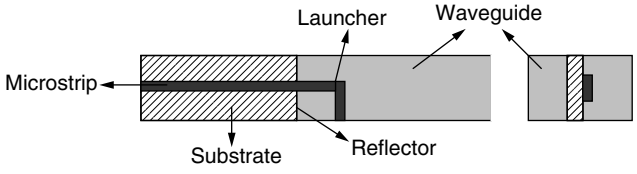


Figure 18. The microstrip end launcher.

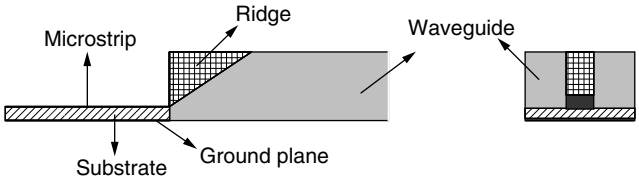


Figure 19. The ridged waveguide transition.

empirically. A possibility to analyze this transition is to use the T-matrix concept of a uniform ridged waveguide section. The advantages and disadvantages of the ridged waveguide transition are similar to those of the finline transition.

3.3.5. Radiating Slot Transition. The radiating slot transition is shown in Fig. 20. A slot in the ground plane of the microstrip is used to excite the waveguide.

In Ref. 22 a theoretical model of the input impedance of the radiating slot transition is given. The model is based on an assumed E-field distribution in the slot and charge distribution on the microstrip. The input impedance is calculated by using the complex power flow through the slot, and the modal voltage discontinuity in the microstrip. Unfortunately, the author does not give experimental results to verify the mathematical models. Nevertheless, simulation results show -20 dB bandwidth over 2% in the X band. Advantages of the transition are its simple structure and the use of the stub as a matching network. Disadvantages are its narrow bandwidth and perpendicular configuration, which is less suitable for miniature arrays.

3.3.6. Tapered Microstrip Transition. Figure 21 shows two views of the tapered microstrip transition, where the tapered-microstrip conductor and the ground plane are connected with the upper and lower waveguide walls, respectively. The waveguide is excited via a slot between the microstrip and the waveguide. Figure 22 shows the field patterns of a microstrip and a waveguide [23]. The microstrip has an E-field distribution that is almost

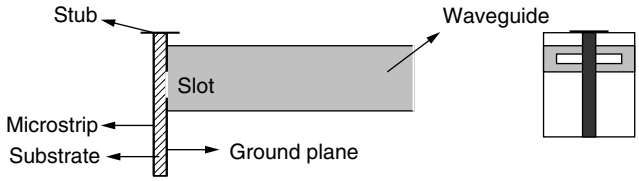


Figure 20. Radiating slot transition.

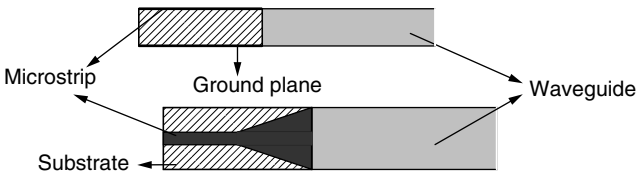


Figure 21. The tapered microstrip transition.

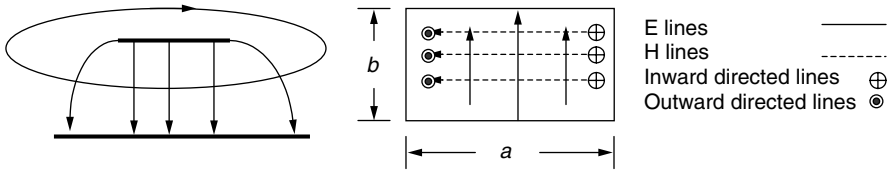


Figure 22. Field patterns of waveguide (at $t = T/4$) and microstrip [23].

uniform between the strip conductor and the ground plane, while the waveguide has a cosinusoidal E-field distribution. In addition, the H-field of the microstrip circle around the strip conductor, while the H-field lines of the waveguide circle around its E-field. It is obvious that a direct transition between these transmission lines will cause severe reflections. Another problem concerning this transition is that the waveguide height must be about the same as the microstrip height (usually less than 1 mm). An advantage of this transition is its longitudinal structure.

3.3.7. Analysis of the MEL. In this section the microstrip end launcher of Fig. 23 is analyzed. It shows a dielectric filled waveguide (DFW) transition where a printed circuit board is placed inside the waveguide. In order to avoid discontinuity effects such as LSM and LSE modes, the waveguide is filled with a dielectric material constant ε_r, which is the same as the dielectric constant of the DFW and the substrate of the microstrip line.

The current loop is divided into two different sections: the z-directed current section, which extends from the plane $z = 0$ to $z = z_1$; and the x-directed section, from $x = 0$ to $x = x_1$. The current is assumed to be continuous at the connecting point $x = x_1$ and $z = z_1$. The perfect ground planes, which are formed by the waveguide walls, are located at $x = 0$ and $x = a$, $y = 0$ and $y = b$, and $z = 0$ (the reflector). The current strip in the plane $y = y_1$ is assumed to be infinitely thin. The width $2w$ is sufficiently narrow so that the current distribution does not vary considerably in the transverse direction. In addition, for simplicity of the analysis, the effects due to the aperture in the reflector are neglected. The efficiency of the transition is characterized by the analysis of the input reflection coefficients.

3.3.8. Reflection Coefficient. The input reflection is given by

$$\Gamma_{\text{in}} = \frac{Z_{\text{in}} - Z_0}{Z_{\text{in}} + Z_0} = S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{22}\Gamma_L} \tag{37}$$

where Z_0 and Z_{in} are the characteristic and input impedance of the microstrip line and the transition, respectively and Γ_L is the reflection coefficient of the load. When the load is not matched, the input reflection can be calculated by using the scattering coefficients $(S_{11}, S_{21}, S_{12}, S_{22})$ of the transition.

3.3.9. Input Impedance. The input impedance seen by the microstrip line satisfies the expression

$$Z_{\text{in}} = -\int_v \frac{E_z \cdot J_z}{I_{\text{in}}^2}\, dV - \int_v \frac{E_x \cdot J_x}{I_{\text{in}}^2}\, dV \tag{38}$$

where E_x and E_z are the electric fields inside the waveguide due to the current density components J_x and J_z, respectively. The current distribution is described as

$$\begin{aligned} J_z &= I_0 \cos[k(z_1 + x_1 - z)]\delta(y - y_1) \\ J_x &= I_0 \cos(kx_1)\delta(y - y_1) \end{aligned} \tag{39}$$

where I_0 is the amplitude of the input current. The current densities J_x and J_z are valid for the region $0 \le z \le z_1$ and $(x_1 - w) \le x \le (x_1 + w)$, and the region $0 \le x \le x_1$ and

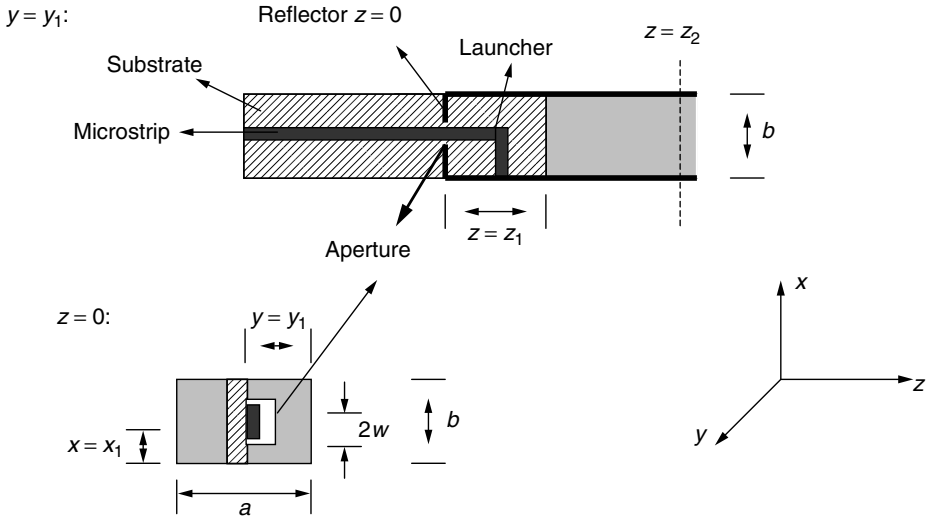


Figure 23. Microstrip line end launcher in a DFW.

$(z_1 - w) \le z \le (z_1 + w)$. The total input current I_{in} at the reference plane $z = 0$ becomes

$$I_{\text{in}} = 2wI_0 \cos[k(z_1 + x_1)] \tag{40}$$

where w is the strip width. Since there is a reflector at the plane $z = 0$, the excited field is caused by the current distribution given by Eq. (35) and by its image

$$\begin{aligned} J_z &= I_0 \cos[k(z_1 + x_1 + z)]\delta(y - y_1) \\ J_x &= -I_0 \cos(kx_1)\delta(y - y_1) \end{aligned} \tag{41}$$

which is valid for the regions $-z_1 \le z \le 0$ and $(x_1 - w) \le x \le (x_1 + w)$, and the region $0 \le x \le x_1$ and $-(z_{1+}w) \le z \le (-z_1 + w)$, respectively. Figure 24 shows the current distribution, its image and the integration domains V and V'.

Figure 25 shows the steps necessary to calculate the input impedance for further analysis. The electric fields are related to the magnetic potentials via [24]

$$\begin{aligned} E_z &= \frac{1}{j\omega\varepsilon}\left(\frac{\partial A_z}{\partial z^2} + k^2 A_z\right) \\ E_x &= \frac{1}{j\omega\varepsilon}\left(\frac{\partial A_x}{\partial x^2} + k^2 A_x\right) \end{aligned} \tag{42}$$

The magnetic vector potentials are defined as

$$\begin{aligned} A_x &= \int_{v'} G_{xx}\left(\frac{x,y,z}{x',y',z'}\right) \cdot J_x(x',y',z')\, dV' \\ A_z &= \int_{v'} G_{zz}\left(\frac{x,y,z}{x',y',z'}\right) \cdot J_z(x',y',z')\, dV' \end{aligned} \tag{43}$$

The primed coordinates x', y', z' represent the source point, while the unprimed coordinates represent the field points.

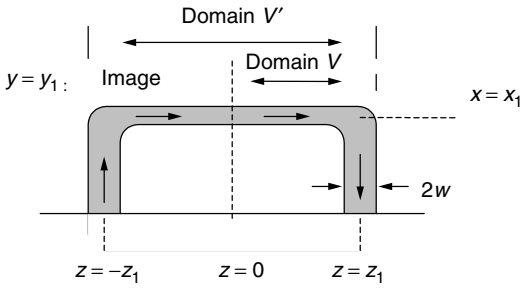


Figure 24. Current distribution and its image of the microstrip end launcher.

The Green function is given by [20]

$$\begin{aligned} G_{xx}\left(\frac{x,y,z}{x',y',z'}\right) &= \sum_{n=0}^{\infty}\sum_{m=1}^{\infty} \frac{\delta_n}{ab\gamma_{mn}} \cos\left(\frac{n\pi x}{b}\right)\sin\left(\frac{m\pi y}{a}\right) \\ &\quad \times \cos\left(\frac{n\pi x'}{b}\right)\sin\left(\frac{n\pi y'}{a}\right) e^{-\gamma_{mn}|z-z'|} \\ G_{zz}\left(\frac{x,y,z}{x',y',z'}\right) &= \sum_{n=1}^{\infty}\sum_{m=1}^{\infty} \frac{1}{2ab\gamma_{mn}} \cos\left(\frac{n\pi x}{b}\right)\sin\left(\frac{m\pi y}{a}\right) \\ &\quad \times \cos\left(\frac{n\pi x'}{b}\right)\sin\left(\frac{n\pi y'}{a}\right) e^{-\gamma_{mn}|z-z'|} \end{aligned} \tag{44}$$

where

$$\gamma_{mn} = \begin{cases} \sqrt{k^2 - \left[\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2\right]} & \text{if } \left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 \le k^2 \\ j\sqrt{\left[\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2\right] - k^2} & \text{otherwise} \end{cases} \tag{45}$$

Substituting (40) in (39), using (37) and performing the integration over V' leads to the following expressions:

$$\begin{aligned} A_x &= \sum_{n=0}^{\infty}\sum_{m=1}^{\infty} \frac{I_0\delta_n}{ab\gamma_{mn}} \cos\left(\frac{n\pi x}{b}\right)\sin\left(\frac{m\pi y}{a}\right) \\ &\quad \times \left[\begin{array}{l} \displaystyle\int_0^a\int_0^b \cos\left(\frac{n\pi x'}{b}\right)\cos(kx')\sin\left(\frac{n\pi y'}{a}\right) \\ \qquad \times \delta(y' - y_1)\, dx'\, dy' \\ \displaystyle\left(-\int_{-z_1}^{0} e^{-\gamma_{mn}|z-z'|} + \int_0^{z_1} e^{-\gamma_{mn}|z-z'|}\right) dz' \end{array}\right] \\ &= \sum_{n=1}^{\infty}\sum_{m=1}^{\infty} \frac{I_0}{2ab\gamma_{mn}} \cos\left(\frac{n\pi x}{b}\right)\sin\left(\frac{m\pi y}{a}\right) \\ &\quad \times \left[\begin{array}{l} \displaystyle\int_0^a\int_0^b \cos\left(\frac{n\pi x'}{b}\right)\cos(kx')\sin\left(\frac{n\pi y'}{a}\right) \\ \qquad \times \delta(y' - y_1)\, dx'\, dy' \\ \displaystyle\left(-\int_{-z_1}^{0} e^{-\gamma_{mn}|z-z'|}\cos[k(z_1 + x_1 + z')] \right. \\ \displaystyle\left. + \int_0^{z_1} e^{-\gamma_{mn}|z-z'|}\cos[k(z_1 + x_1 + z')]\right) dz' \end{array}\right] \end{aligned} \tag{46}$$

The integration is performed in the paper by Ho and Shin [20] and is not included here. Inserting the results in (38) leads to the expressions for E_x and E_z. Substituting

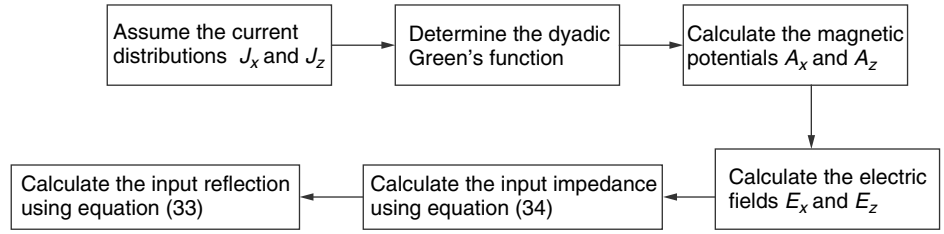


Figure 25. Steps for calculating the input reflection.

E_x and E_z in (34) gives the result for the Z_{in} [20]. It can be shown that

$$
\begin{aligned}
-\int_v \frac{E_z \cdot J_z}{I_{\text{in}}^2}\,dV &= j\sum_{n=1}^{\infty}\sum_{m=1}^{\infty} \frac{1}{\omega\varepsilon_0 ab\gamma_{mn}\cos^2[k(z_1+x_1)]}\frac{b}{n\pi w} \\
&\quad \times \sin\left(\frac{n\pi w}{b}\right)\sin^2\left(\frac{m\pi y_1}{a}\right)\sin^2\left(\frac{n\pi x_1}{b}\right) \\
&\quad \times \left\{ \begin{array}{l} \dfrac{k\sin(k(z_1+x_1))}{k^2+\gamma_{mn}^2}[\gamma_{mn}\cos^2(k(z_1+x_1)) \\ +k\sin(k(z_1+x_1)) - e^{-\gamma_{mn}z_1}(\gamma_{mn}\cos(kx_1) \\ +k\sin(kx_1))] \\ -\left(\begin{array}{c} \dfrac{e^{-\gamma_{mn}z_1}(\gamma_{mn}\cos(kx_1)+k\sin(kx_1))}{k^2+\gamma_{mn}^2} \\ \times k\sin(k(z_1+x_1)) \\ +\gamma_{mn}\cos(kx_1)\sin h(\gamma_{mn}z_1) \\ -k\sin(kx_1)\cos h(\gamma_{mn}z_1) \end{array} \right) \end{array} \right\} \\
-\int_v \frac{E_x \cdot J_x}{I_{\text{in}}^2}\,dV &= j\sum_{n=0}^{\infty}\sum_{m=1}^{\infty} \frac{240\delta_n}{abk\gamma_{mn}}\frac{\sin h(\gamma_{mn}w)}{\gamma_{mn}w} \\
&\quad \times \sin^2\left(\frac{m\pi y_1}{a}\right)\sin h(\gamma_{mn}z_1) \\
&\quad \times e^{-\gamma_{mn}z_1}\frac{\left[\sin(kx_1)\cos\left(\dfrac{n\pi x_1}{b}\right) - \dfrac{n\pi}{ak}\cos(kx_1)\sin\left(\dfrac{n\pi x_1}{b}\right)\right]^2}{\cos^2(k(z_1+x_1))\left(1-\left(\dfrac{n\pi}{ak}\right)^2\right)}
\end{aligned}
\tag{47}
$$

The microstrip end launcher may be used to excite a DFW with two E-plane steps. This kind of waveguide will be discussed in Section 6. The waveguide with E-plane steps (Fig. 26) is filled with a dielectric material with a dielectric constant $\varepsilon_r = 2.53$. The airgap matching network is discussed in Section 5.

The effect of different parameters such as x_1, y_1, and z_1 (see Fig. 23) on the behavior of the input impedance in the X and Ka bands has been studied by Ho and Shih [20] and Lam et al. [25], respectively. An optimization routine has been developed in order to obtain the optimal values for the different parameters. Since the purpose is to realize a miniature antenna with a large bandwidth, a two $\lambda/4$ matching network is employed. Figure 27 illustrates this concept. The design parameters x_1, y_1, z_1, Z_{0k} and l are optimized. Several parameters were kept constant throughout the analysis, such as the physical dimensions of the width w and the height h of the microstrip, and the width of the DFW with steps. The values of w, a and b_3 were chosen to be 0.185, 14, and 5 mm, respectively. A substrate 0.25 mm thick with a dielectric constant of 2.53 is used as a dielectric slab. The end launcher is matched to a microstrip line with a characteristic impedance of 75 Ω.

Figure 28 shows the optimized calculated input reflection with and without a $\lambda/4$ section. It is observed that the $\lambda/4$ section has a remarkable effect on the bandwidth. The −20 dB bandwidth increases by almost 15%. The difference is due to the fact that the $\lambda/4$ section decreases the frequency sensitivity of the multiple reflections [23].

3.3.10. Design of the Microstrip End Launcher. Figure 29 shows the geometry of the end launcher with two $\lambda/4$ sections and its integration into the DFW. The thin substrate consists of the microstrip line with z_{01}. Two $\lambda/4$ sections, z_{0j} and z_{0k}, and the current loop launcher were inserted into the DFW.

In order to obtain a sufficiently narrow loop microstrip, a thin microstrip substrate with a dielectric constant of $\varepsilon_r = 2.53$ and $h = 0.25$ mm is chosen. This thickness was chosen because the dielectric constant of the substrate is not exactly the same as the dielectric constant of DFW. Furthermore, a thin substrate would cause fewer LSM and LSE mode effects.

The $\lambda/4$ section with z_{0j} is chosen to have the same width as the end launcher loop. The width of the $\lambda/4$ section with z_{0k} has been optimized, $w_1 = 0.32$ mm. The

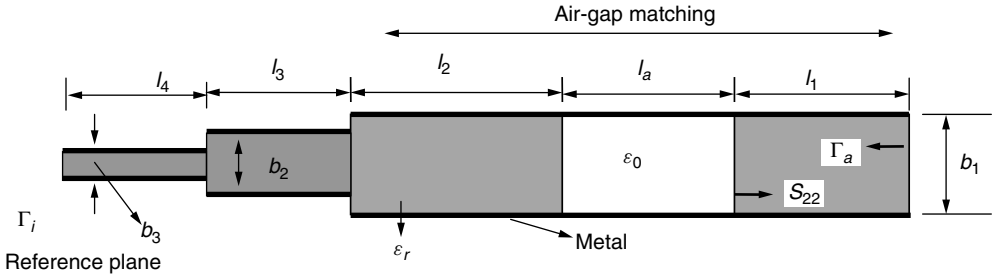


Figure 26. Two-plane step DFW.

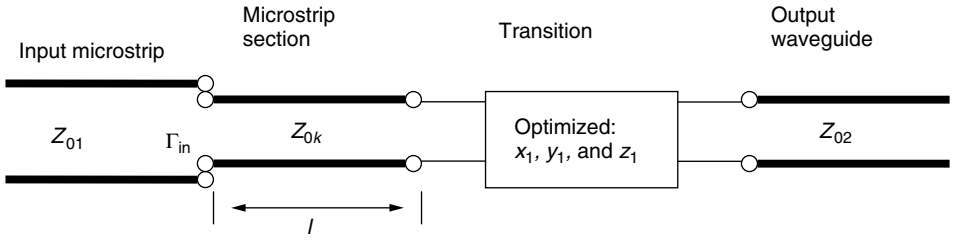


Figure 27. Optimization of matching network.

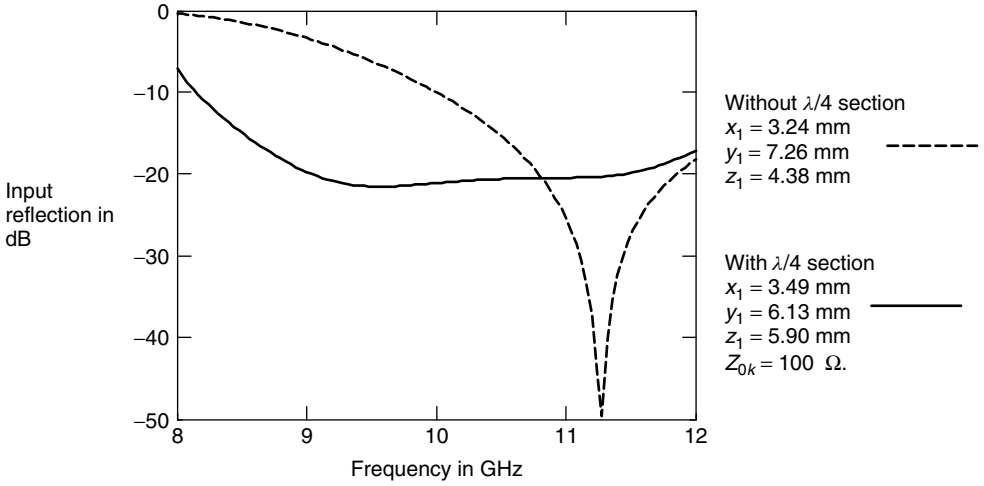


Figure 28. Calculated input reflection of the microstrip end launcher with and without $\lambda/4$ section.

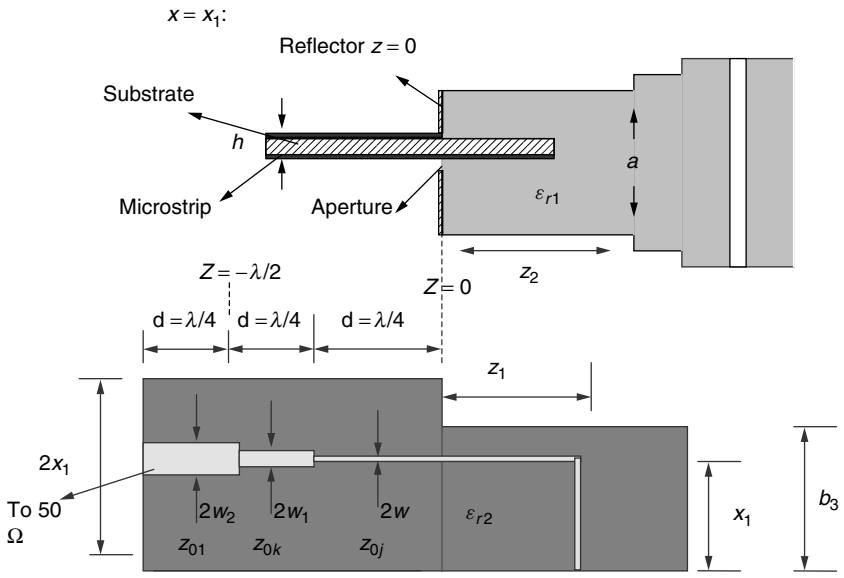


Figure 29. Microstrip end launcher transition at 10 GHz and its integration to DFW. (Courtesy of IRCTR.)

corresponding microstrip width is $w_2 = 0.37$ mm at the design frequency of 10 GHz. The characteristic impedance of the input microstrip line Z_{01} is chosen to be 75 Ω and has the same length as the $\lambda/4$ section. The current in the loop is assumed to be continuous. Therefore, a bend having a radius of $r = 4w$ is used to minimize the reflections [23] (not shown in Fig. 29).

Figure 30 shows the optimized design parameters corresponding to those in Fig. 29. Figure 31 shows the calculated input reflection as a function of frequency at different reference planes of the microstrip end launcher. The reference planes are $z = z_2$, $z = 0$, and $z = -\lambda/2$ (see Fig. 29). These planes indicate the input reflection of the DFW [26], the transition of the microstrip end launcher with DFW, and the total input reflection of the microstrip end launcher with the DFW and the two $\lambda/4$ sections.

The input reflection of the DFW has a −20 dB bandwidth over a 15% frequency band. There are two dips at 9.1 and 10.3 GHz. The input reflection of MEL with

a	14 mm		
b_3	5 mm		
$\varepsilon_{r1} = \varepsilon_{r1}$	2.33		
h	0.25 mm		
x_1	3.9 mm		
y_1	9.48 mm		
z_1	6.06 mm		
z_2	50 mm		
Z_{oj}	75 ohm	w	0.185 mm
Z_{0k}	55 ohm	w_1	0.32 mm
Z_{01}	50 ohm	w_2	0.37 mm

Figure 30. Optimal design parameters.

DFW has a −20 dB bandwidth over 5% at the frequency band. In addition, the dips at 9.1, 9.8, and 10.6 GHz can be distinguished.

A piece of the dielectric material inside the waveguide was removed in order to place the launcher section

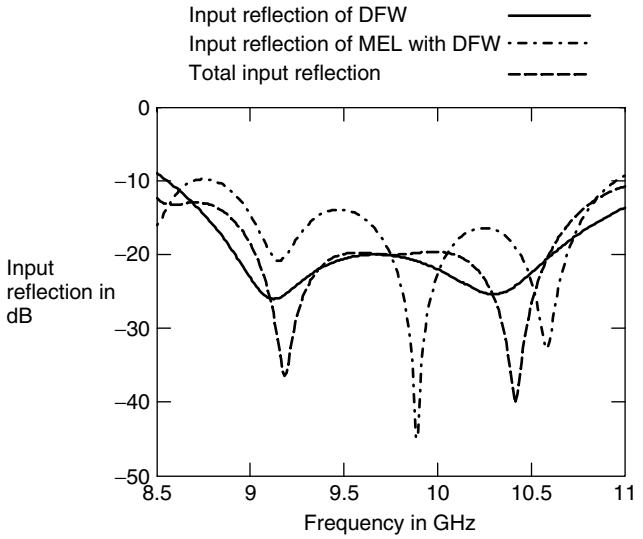


Figure 31. Input reflections as function of frequency at different reference planes of MEL.

of the microstrip circuit inside the waveguide. Then, an electrical connection between the launcher and the lower waveguide wall was made. A reflector was placed to close the waveguide, and practically all waves are reflected in the desired direction. To prevent a short circuit, a small aperture was made in the reflector wall. Since the theoretical model does not take the aperture effect into account, the dimensions of the aperture were chosen experimentally. A conducting post was used for the electrical connection between the end launcher and the waveguide wall [25].

The coax-to-microstrip transition has been realized empirically by tapering the pin of the SMA connector. The half-circle of the mounting plate with radius D is used to compensate the reactance of the coax-to-microstrip transition empirically [25,27]. Measurement results can be seen in Fig. 32.

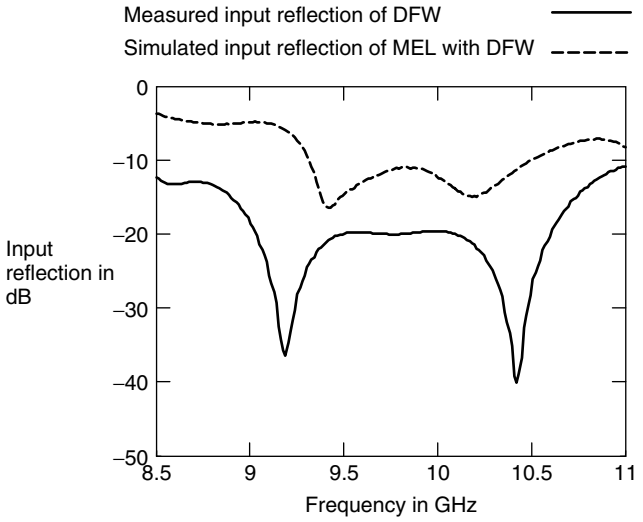


Figure 32. Measured and simulated results of input reflection of MEL as a function of frequency.

4. ATTENUATION

Although the waveguide has low losses, the electromagnetic waves still suffer from attenuation. Since the waveguide metal is not a perfect conductor, there are some ohmic (conduction) losses. If the waveguide is filled with dielectric material, an extra factor, called the *dielectric losses*, which contributes to the total losses, needs to be taken into the consideration. They are denoted with α_c and α_d, respectively, and are given by the following expression [11]

$$\alpha_c = \frac{2R_s}{\delta_m \delta_n b \eta \sqrt{1-\left(\frac{f_c}{f}\right)^2}} \left\{ \left(\delta_m + \delta_n \frac{b}{a}\right)\left(\frac{f_c}{f}\right)^2 + \frac{b}{a}\left[1-\left(\frac{f_c}{f}\right)^2\right] \frac{m^2 ab + (na)^2}{(ma)^2 + (na)^2} \right\} \tag{48}$$

where

$$R_s = \sqrt{\frac{\omega\mu}{2\sigma}} \quad \text{for } \sigma \gg \omega\varepsilon$$
$$\delta_m = \begin{cases} 2 & m = 0 \\ 1 & m \neq 0 \end{cases} \tag{49}$$

and

$$\alpha_d = 8.68 \left(\frac{\varepsilon_e''}{\varepsilon_e'}\right) \frac{\pi}{\lambda} \left(\frac{\lambda_g}{\lambda}\right) \quad \text{dB/m} \tag{50}$$

Figure 33 shows the TE_{10} conduction losses as a function of frequency for three different dielectric materials.

5. MINIATURIZATION TECHNIQUE

Large array antennas can benefit significantly from miniaturization. Since the propagation of the fundamental mode in the rectangular waveguide is independent of the height, the miniaturization can be achieved by lowering the height. Filling the waveguide with dielectric material can also contribute to miniaturization. When a dielectric

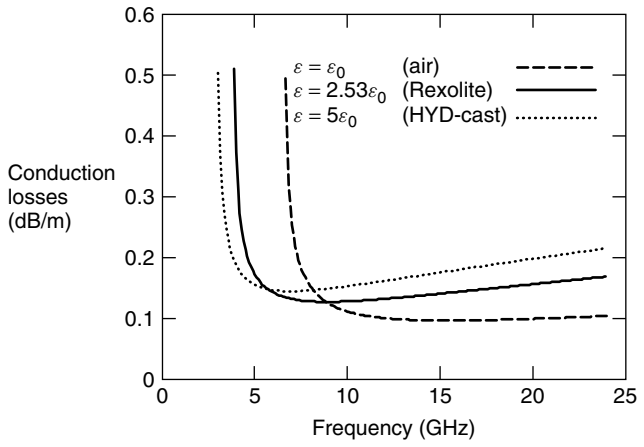


Figure 33. The conduction losses as function of frequency for different materials. (*Source*: C. A. Balanis, *Advanced Engineering Electromagnetics*, Wiley, 1989.)

is used, the dimensions of the antenna will decrease by the square root of the dielectric constant.

This dielectric loading technique is also applied to protect the waveguide from environmental conditions and makes it possible to flesh mount the antenna on the surface of the spacecraft or aircraft. Figure 34 shows a number of dual-polarized DFW in different frequency bands.

Using this technique for miniaturization may in some cases lead to a high aperture reflection. This leads to a complexity in aperture matching. In this section the aperture matching technique is presented.

5.1. Aperture Characteristics

In order to derive an expression for the aperture admittance, the waveguide geometry in Fig. 35 is considered. The approach is based on power conservation across the aperture from region I to region II. It is assumed that the aperture is mounted in a perfectly conducting plane of infinite size. It is also assumed that the excitation is such that only symmetrical TE_{m0} modes are generated in region I (in Fig. 35).

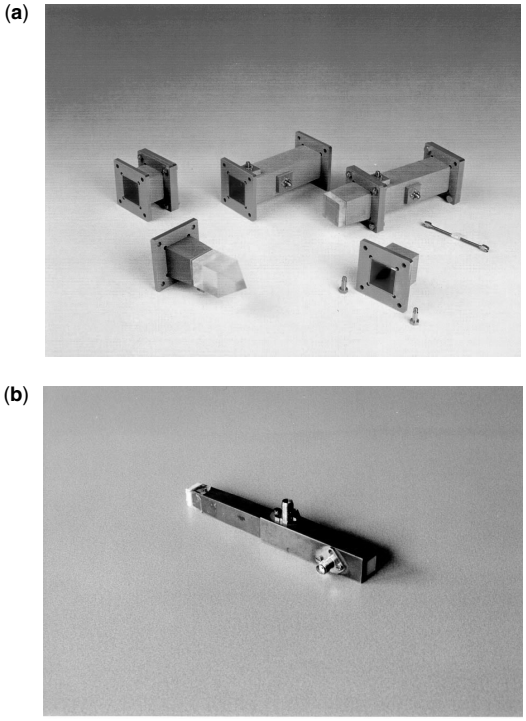


Figure 34. Miniaturized dual-polarized DFW: (**a**) S band, $\varepsilon_r = 2.53$; (**b**) X band, $\varepsilon_r = 5.0$. (Courtesy of IRCTR.)

5.1.1. Internal Field in Region I. When only the TE_{m0} mode (may) exist in the waveguide, the field components in region I (Fig. 35) can be written as [28]

$$\begin{aligned} E_y^{\mathrm{I}} &= E_0(e^{-j\beta_{z1}z} + \Gamma e^{j\beta_{z1}z})\cos\left(\frac{\pi x}{a}\right) \\ &\quad + \sum_{m=3,5,\ldots}^{\infty} A_m \cos\left(\frac{m\pi x}{a}\right) e^{j\beta_{zm}z} \\ E_x^{\mathrm{I}} &= H_y^{\mathrm{I}} = 0 \\ H_x^{\mathrm{I}} &= Y_{10}E_0(e^{-j\beta_{z1}z} - \Gamma e^{j\beta_{z1}z})\cos\left(\frac{\pi x}{a}\right) \\ &\quad - \sum_{m=3,5,\ldots}^{\infty} Y_{m0}A_m \cos\left(\frac{m\pi x}{a}\right) e^{j\beta_{zm}z} \end{aligned} \tag{51}$$

where Y_0 is the free-space admittance and Y_{m0} the wave impedance for the TE_{m0} modes. In this section the following changes in parameters are introduced:

$$\beta_{z1} = \beta\frac{Y_{10}}{Y_0}, \beta_{zm} = \beta\frac{Y_{m0}}{Y_0}, \beta = \beta\sqrt{\varepsilon_r} \tag{52}$$

Note that in this case the wavenumber is higher than the free-space wavenumber. At the boundary $z = 0$ and with $A_m = D_m E_0(1 + \Gamma)$, Eq. (47) becomes

$$\begin{aligned} E_y^{\mathrm{I}}(x, y, z = 0) &= E_0(1 + \Gamma)\left(\cos\left(\frac{\pi x}{a}\right)\right. \\ &\quad \left. + \sum_{m=3,5,\ldots}^{\infty} D_m \cos\left(\frac{m\pi x}{a}\right)\right) \\ E_x^{\mathrm{I}}(x, y, z = 0) &= H_y^{\mathrm{I}}(x, y, z = 0) = 0 \\ H_x^{\mathrm{I}}(x, y, z = 0) &= Y_{10}E_0(1 - \Gamma)\cos\left(\frac{\pi x}{a}\right) \\ &\quad - \sum_{m=3,5,\ldots}^{\infty} Y_{m0}D_m E_0(1 + \Gamma)\cos\left(\frac{m\pi x}{a}\right) \end{aligned} \tag{53}$$

The expression for the unknown coefficients D_m and the normalized aperture admittance will now be derived.

5.1.2. Aperture Admittance. The reaction integral I for the aperture fields are considered to compute the energy transfer through the aperture. The integral for region I becomes

$$I = \int_{-\frac{a}{2}}^{\frac{a}{2}} \int_{-\frac{b}{2}}^{\frac{b}{2}} E_y^{\mathrm{I}}(x, y, 0) H_x^{\mathrm{I}}(x, y, 0)\, dy\, dx \tag{54}$$

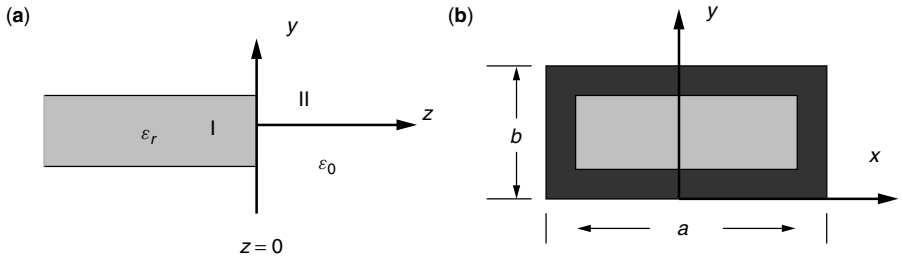


Figure 35. Configuration of the waveguide for analysis of aperture admittance: (**a**) side view; (**b**) front view [28].

Substituting Eq. (49) in (50) and performing the integration leads to

$$I = \frac{ab}{2} Y_{10} E_0^2 (1+\Gamma)^2 \left[\frac{1-\Gamma}{1+\Gamma} - \sum_{m=3,5,\ldots}^{\infty} \frac{Y_{m0}}{Y_{10}} D_m^2 \right] \tag{55}$$

Rearranging Eq. (51) gives the following expression for the normalized aperture admittance:

$$y_{\text{ap}} = \frac{1-\Gamma}{1+\Gamma} = \frac{2}{ab} \frac{I}{Y_{10} E_0^2 (1+\Gamma)^2} + \sum_{m=3,5,\ldots}^{\infty} \frac{Y_{m0}}{Y_{10}} D_m^2 \tag{56}$$

In order to calculate the integral I, it is assumed that the tangential fields are continuous across the aperture. The following relationship should exist:

$$\begin{aligned} I &= \int_{-(a/2)}^{a/2} \int_{-(b/2)}^{b/2} E_y^I(x,y,0) H_x^{\text{I}}(x,y,0)\, dy\, dx \\ &= \int_{-(a/2)}^{a/2} \int_{-(b/2)}^{b/2} E_y^{\text{II}}(x,y,0) H_x^{\text{II}}(x,y,0)\, dy\, dx \end{aligned} \tag{57}$$

In region II the aperture fields can be expressed in the spectral domain via

$$\hat{E}_y(k_x,k_y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} E_y(x,y,0) e^{j\{k_x x + k_y y\}}\, dy\, dx \tag{58}$$

where $\hat{E}_y(k_x,k_y)$ is the Fourier transformation of the aperture electric field and k_x, k_y are the spectral frequencies that extend over the entire frequency spectrum $-\infty \le k_x, k_y \le \infty$. Since the tangential fields are zero outside the aperture and are even functions with respect to x and y, Parseval's theorem can be used to obtain the following relationship [29]:

$$\begin{aligned} I &= \int_{-(a/2)}^{a/2} \int_{-(b/2)}^{b/2} E_y^{\text{II}}(x,y,0) H_x^{\text{II}}(x,y,0)\, dy\, dx \\ &= \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \hat{E}_y^{\text{II}}(k_x,k_y,0) \hat{H}_y^{II}(k_x,k_y,0)\, dk_y\, dk_x \end{aligned} \tag{59}$$

Substituting (55) in (52) relates the normalized aperture admittance to the exterior fields in the spectral domain. The result becomes

$$\begin{aligned} y_{\text{ap}} = \frac{1-\Gamma}{1+\Gamma} &= \frac{2}{abY_{10}E_0^2(1+\Gamma)^2} \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \hat{E}_y^{\text{II}}(k_x,k_y,0) \\ &\quad \times \hat{H}_y^{\text{II}}(k_x,k_y,0)\, dk_y\, dk_x + \sum_{m=3,5,\ldots}^{\infty} \frac{Y_{m0}}{Y_{10}} D_m^2 \end{aligned} \tag{60}$$

The next step is to solve the fields in the exterior region.

5.1.3. Exterior Field in Region II. The expressions for the exterior field are derived using the electric and magnetic potentials. In terms of the potentials, the electromagnetic fields are given as [28]

$$\begin{aligned} E &= -\nabla \times F - j\omega A + \frac{\nabla\nabla \cdot A}{j\omega\varepsilon} \\ H &= \nabla \times A - j\omega F + \frac{\nabla\nabla \cdot F}{j\omega\mu} \end{aligned} \tag{61}$$

where $\nabla \cdot A$ is the divergence of A and ∇ is the gradient operator [12]. The field components in region II can be expressed as [29]

$$\begin{aligned} E_y^{\text{II}} &= -\frac{\partial \psi}{\partial x} + \frac{1}{j\omega\varepsilon} \frac{\partial^2 \varphi}{\partial y\, \partial z} \\ E_x^{\text{II}} &= \frac{\partial \psi}{\partial y} + \frac{1}{j\omega\varepsilon} \frac{\partial^2 \varphi}{\partial x\, \partial z} \\ H_y^{\text{II}} &= \frac{1}{j\omega\mu} \frac{\partial^2 \psi}{\partial y\, \partial z} + \frac{\partial \varphi}{\partial x} \\ H_x^{\text{II}} &= \frac{1}{j\omega\mu} \frac{\partial^2 \psi}{\partial x\, \partial z} - \frac{\partial \varphi}{\partial y} \end{aligned} \tag{62}$$

where $F = \psi\hat{z}$, $A = \varphi\hat{z}$. A possible solution of Eq. (58) in the spectral domain is given by

$$\begin{aligned} \psi &= \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(k_x,k_y) e^{-j(k_x x + k_y y + k_z z)}\, dk_y\, dk_x \\ \varphi &= \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(k_x,k_y) e^{-j(k_x x + k_y y + k_z z)}\, dk_y\, dk_x \end{aligned} \tag{63}$$

The modal coefficients f and g are derived as follows. The Fourier transform of the fields in region II is given as

$$\begin{aligned} [E^{\text{II}}(x,y), H^{\text{II}}(x,y)] = \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} &[\hat{E}^{\text{II}}(k_x,k_y), \\ &\times \hat{H}^{\text{II}}(k_x,k_y)] e^{-j\{k_x x + k_y y\}}\, dk_y\, dk_x \end{aligned} \tag{64}$$

Substituting Eqs. (59) and (60) into (58) leads to

$$\begin{aligned} \hat{E}_y^{\text{II}}(k_x,k_y) &= j\left(k_x f + \frac{k_x k_z}{\omega\varepsilon} g \right) e^{-jk_z z} \\ \hat{E}_x^{\text{II}}(k_x,k_y) &= -j\left(k_y f - \frac{k_x k_z}{\omega\varepsilon} g \right) e^{-jk_z z} \end{aligned} \tag{65}$$

The tangential electric field components are continuous across the aperture at $z = 0$:

$$\begin{aligned} \hat{E}_y^{\text{II}}(k_x,k_y,0) &= \hat{E}_y^{\text{I}}(k_x,k_y,0) \\ \hat{E}_x^{\text{II}}(k_x,k_y,0) &= \hat{E}_x^{\text{I}}(k_x,k_y,0) \end{aligned} \tag{66}$$

The modal coefficients f and g can be expressed in terms of the electric field components in region I:

$$\begin{aligned} f(k_x,k_y) &= j \frac{k_y \hat{E}_x^{\text{I}}(k_x,k_y,0) - k_x \hat{E}_y^{\text{I}}(k_x,k_y,0)}{k_y^2 + k_x^2} \\ g(k_x,k_y) &= -j \frac{(k_y \hat{E}_x^{\text{I}}(k_x,k_y,0) + k_x \hat{E}_y^{\text{I}}(k_x,k_y,0))\omega\varepsilon}{k_z(k_y^2 + k_x^2)} \end{aligned} \tag{67}$$

From Eqs. (58), (60), and (63) it can be deduced that

$$\hat{H}_x^{\text{II}}(k_x,k_y,0) = \frac{(k^2 - k_x^2)\hat{E}_x^{\text{I}}(k_x,k_y,0) + k_x k_y \hat{E}_y^{\text{I}}(k_x,k_y,0)}{\omega\mu k_z} \tag{68}$$

Inserting the Fourier transform of the electric field components from Eq. (49) yields

$$\begin{aligned}\hat{E}_x^{\mathrm{II}}(k_x, k_y, 0) &= 0\\ \hat{E}_y^{\mathrm{II}}(k_x, k_y, 0) &= \hat{E}_y^{\mathrm{I}}(k_x, k_y, 0)\\ &= E_0(1+\Gamma)\iint_{\mathrm{ap}}\left(\cos\left(\frac{\pi x}{a}\right)\right.\\ &\quad\left.+\sum_{m=3,5,\ldots}^{\infty} D_m\cos\left(\frac{m\pi x}{a}\right)\right)e^{+j\{k_x x+k_y y\}}\,dy\,dx\end{aligned} \tag{69}$$

Substituting $\hat{E}_y^{\mathrm{II}}(k_x, k_y, 0)\hat{H}_y^{\mathrm{II}}(k_x, k_y, 0) = \hat{E}_y^{\mathrm{I}}(k_x, k_y, 0)\hat{H}_y^{\mathrm{I}}(k_x, k_y, 0)$ in Eq. (57) gives the following scheme for the aperture admittance

$$\begin{aligned}y_{\mathrm{ap}} &= \frac{2}{abY_{10}\omega\mu}\frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{(k^2-k_x^2)}{k_z}C_0^2(k_x)\\ &\quad\times\left[C_1^2(k_x)+2C_1(k_x)\sum_{m=3,5,\ldots}^{\infty}D_mC_m(k_x)\right.\\ &\quad\left.+\left(\sum_{m=3,5,\ldots}^{\infty}D_mC_m(k_x)\right)^2\right]dk_y\,dk_x+\sum_{m=3,5,\ldots}^{\infty}\frac{Y_{m0}}{Y_{10}}D_m^2\end{aligned} \tag{70}$$

with

$$\begin{aligned}Y_{ij} = Y_{ji} &= \frac{2}{ab\omega\mu}\frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{(k^2-k_x^2)}{k_z}\\ &\quad\times C_0^2(k_y)C_j(k_x)C_j(k_x)\,dk_y\,dk_x\end{aligned} \tag{71}$$

for $i, j \neq 0$ the expression results in

$$y_{\mathrm{ap}} = \frac{Y_{11}}{Y_{10}} + 2\sum_{m=3,5,\ldots}D_m\frac{Y_{1m}}{Y_{10}} + \sum_{m=3,5,\ldots}D_m^2\left(\frac{Y_{mm}}{Y_{10}}+\frac{Y_{m0}}{Y_{10}}\right) \tag{72}$$

where

$$\begin{aligned}C_0(k_y) &= \frac{b\sin\left(k_y\frac{b}{2}\right)}{k_y\frac{b}{2}}\\ C_1(k_x) &= \frac{2\pi a\cos\left(k_x\frac{a}{2}\right)}{(\pi)^2-(k_xa)^2}\\ C_m(k_x) &= \frac{2m\pi aj^{m-1}\cos\left(k_x\frac{a}{2}\right)}{(m\pi)^2-(k_xa)^2}\end{aligned} \tag{73}$$

Figure 36 shows the normalized aperture impedance of a DFW as function of frequency in the L and X bands. The height of the waveguide is given as a parameter. The width

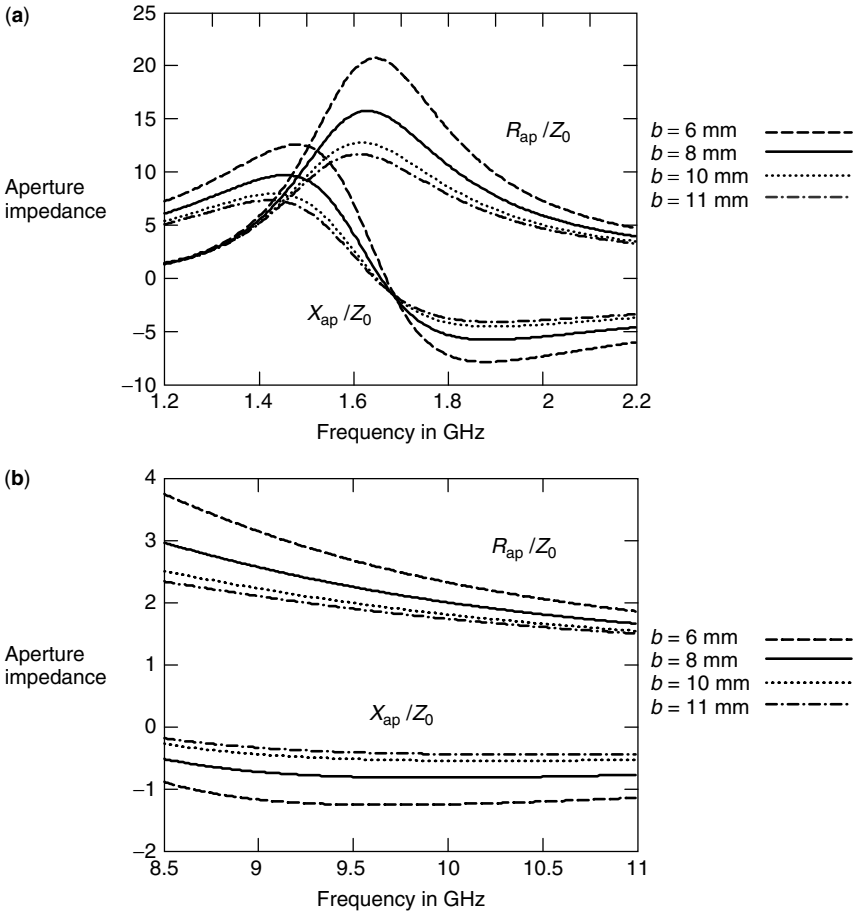


Figure 36. The normalized aperture impedance of DFW as a function of frequency: (**a**) L band; (**b**) X band, $\varepsilon_r = 2.53$.

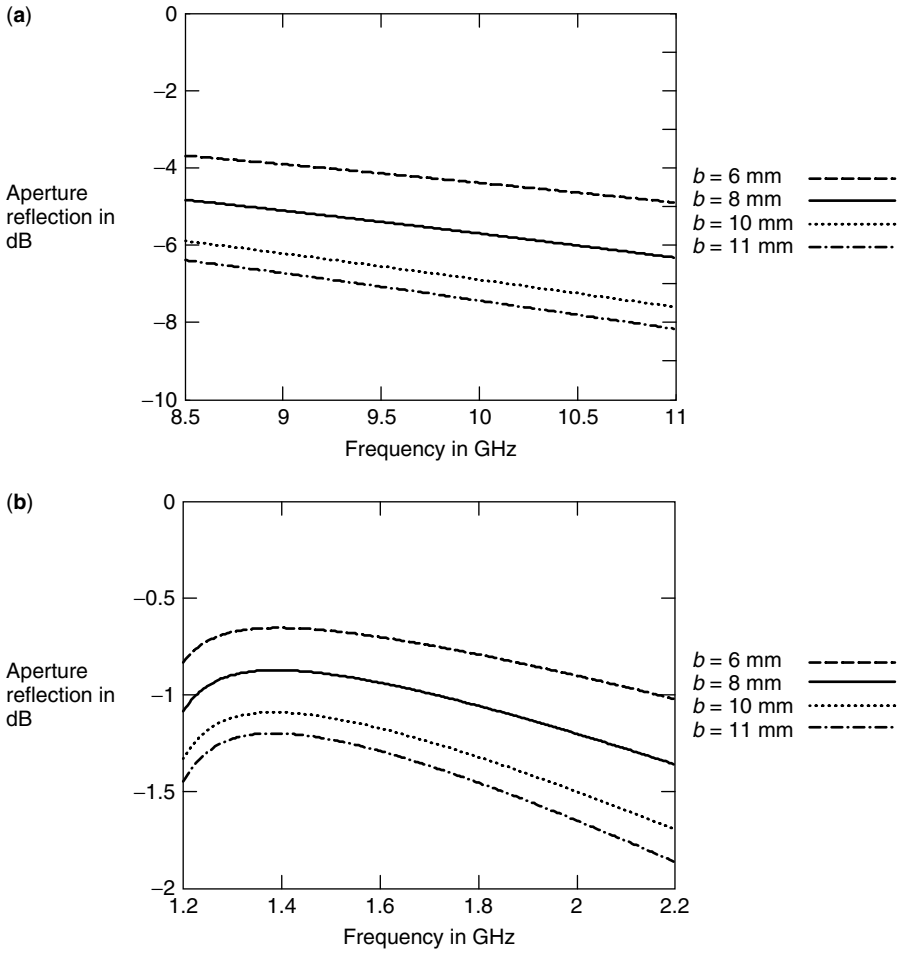


Figure 37. The aperture reflection as a function of frequency: (**a**) L band; (**b**) X band, $\varepsilon_r = 2.53$.

of the aperture is 83 and 17 mm in the L and X bands, respectively.

Figure 37 shows the aperture reflection for different heights of the aperture. The reflection losses increase with the decrease in the height of the aperture.

5.2. Matching Technique

The aperture reflection results given in the previous section indicate that the aperture reflection can be high. It is possible to match the aperture reflection using a microwave matching technique. In Ref. 30 a unique matching technique is proposed and analyzed. The aim of this section is to derive the mathematical expression for such a matching condition. For this aim the matching network is considered as a two-port device. This is shown in Fig. 38.

With the definition for the reflection coefficient

$$\Gamma_i \triangleq \frac{b_1}{a_1}, \Gamma_a \triangleq \frac{a_2}{b_2} \tag{74}$$

where an optimal matching network must satisfy the condition

$$\Gamma_i = 0 \tag{75}$$

The scattering matrix of the two-port network is given as

$$\begin{aligned} b_1 &= S_{11}a_1 + S_{12}a_2 \\ b_2 &= S_{21}a_1 + S_{22}a_2 \end{aligned} \tag{76}$$

Substituting (70) in (72) leads to

$$\begin{aligned} b_1 &= S_{11}a_1 + S_{12}\Gamma_a b_2 \\ b_2 &= S_{21}a_1 + S_{22}\Gamma_a b_2 \end{aligned} \tag{77}$$

Equation (73) leads to

$$\frac{b_1}{a_1} = \frac{S_{11} - (S_{11}S_{22} - S_{21}S_{12})\Gamma_a}{1 - S_{22}\Gamma_a} = \frac{S_{11} - (\det S)\Gamma_a}{1 - S_{22}\Gamma_a} \tag{78}$$

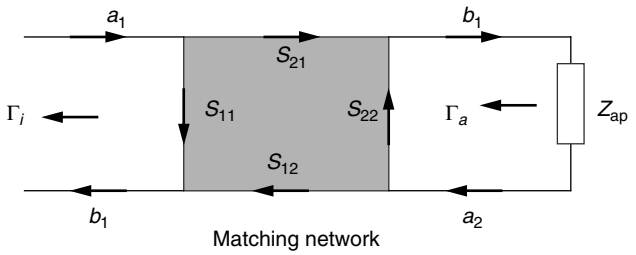


Figure 38. The two-port matching network and its scattering diagram [30].

Equation (74) can be rearranged into

$$\Gamma_i = \frac{b_1}{a_1} = \det S \frac{\dfrac{S_{11}}{\det S} - \Gamma_a}{1 - S_{22}\Gamma_a} \tag{79}$$

when considering

$$S_{22}^* \det S = S_{22}^*(S_{11}S_{22} - S_{21}S_{12}) \tag{80}$$

It has been shown [29] that a symmetric and lossless two-port device has the properties

$$\begin{aligned} S_{11}S_{22}^* &= -S_{12}S_{22}^* \\ |S_{11}| &= |S_{22}| \\ |S_{12}| &= \sqrt{1 - |S_{11}|^2} \end{aligned} \tag{81}$$

The S parameters at an arbitrary reference plane can be characterized by

$$\begin{aligned} &S_{11} = |S_{11}|e^{j\theta_1}, \quad S_{22} = |S_{22}|e^{j\theta_2} = |S_{11}|e^{j\theta_2} \\ &S_{12} = |S_{12}|e^{j\phi} = \sqrt{1 - |S_{11}|^2}e^{j\phi} = S_{21} \\ &\phi = \frac{\theta_1 + \theta_2}{2} + \frac{\pi}{2} \mp 2n\pi \end{aligned} \tag{82}$$

Equation (76) can now be written as

$$\begin{aligned} \det S S_{22}^* &= S_{11}|S_{11}|^2 - |S_{12}|^2 e^{2j\phi}|S_{11}|e^{-j\theta_2} \\ &= S_{11}|S_{11}|^2 - |S_{12}|^2 e^{2j\phi}|S_{11}|e^{-j\theta_1}e^{j\theta_1}e^{-j\theta_2} \\ &= S_{11}|S_{11}|^2 - |S_{12}|^2 e^{2j\phi}S_{11}e^{-j(\theta_1+\theta_2)} \\ &= S_{11}\{|S_{11}|^2 - (1 - |S_{11}|^2)e^{2j\phi}e^{-j(\theta_1+\theta_2)}\} \\ &= S_{11}\{|S_{11}|^2 + (1 - |S_{11}|^2)e^{2j\phi}e^{-j(\theta_1+\theta_2)}e^{j\pi}\} \end{aligned} \tag{83}$$

or

$$\det S S_{22}^* = S_{11}\{|S_{11}|^2 + (1 - |S_{11}|^2)\} = S_{11} \tag{84}$$

Substituting (80) into (75) gives

$$\Gamma_i = \frac{b_1}{a_1} = \det S \frac{S_{22}^* - \Gamma_a}{1 - S_{22}\Gamma_a} \tag{85}$$

The input reflection given by Eq. (71) can be minimized if S_{22}^* is adjusted to cancel the aperture reflection in amplitude and phase. The following necessary condition must exist for the two-port matching network.

$$\Gamma_a = S_{22}^* \tag{86}$$

5.2.1. Airgap Matching Network. On the basis of condition (82) in this section, it is shown that S_{22}^* can be tuned using an airgap matching network to minimize the input reflection. Consider therefore an air-filled homogeneous waveguide section with finite length l, which is bounded by two waveguide sections filled with dielectric material as illustrated in Fig. 39.

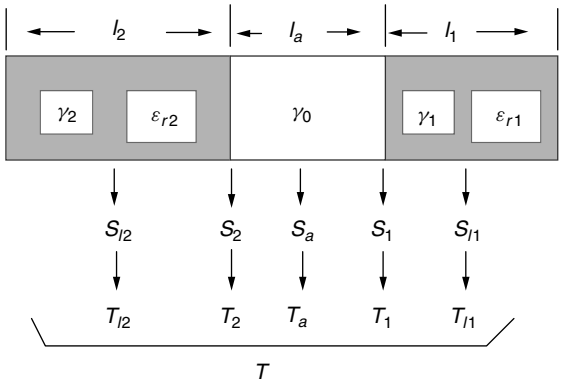


Figure 39. Configuration of S–T matrix with the airgap [30].

The overall T-matrix is formed by the multiplication of a series of successive T matrices

$$T = T_{l2}T_2T_{la}T_1T_{l1} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \tag{87}$$

The scattering matrix S^0 of the two-port networks is related to the transmission matrix T^0 and vice versa as

$$T^0 = \begin{bmatrix} T_{11}^0 & T_{12}^0 \\ T_{21}^0 & T_{22}^0 \end{bmatrix} = \begin{bmatrix} -\dfrac{\Delta S^0}{S_{21}^0} & \dfrac{S_{11}^0}{S_{21}^0} \\ -\dfrac{S_{22}^0}{S_{21}^0} & \dfrac{1}{S_{21}^0} \end{bmatrix} \tag{88}$$

where $\Delta S^0 \triangleq S_{11}^0 S_{22}^0 - S_{21}^0 S_{12}^0$:

$$S^0 = \begin{bmatrix} S_{11}^0 & S_{12}^0 \\ S_{21}^0 & S_{22}^0 \end{bmatrix} = \begin{bmatrix} \dfrac{T_{12}^0}{T_{22}^0} & \dfrac{\Delta T^0}{T_{22}^0} \\ \dfrac{1}{T_{22}^0} & \dfrac{T_{21}^0}{T_{22}^0} \end{bmatrix} \tag{89}$$

From (85) S_{22} is given by

$$S_{22}^0 = -\frac{T_{21}^0}{T_{22}^0} \tag{90}$$

The scattering matrix of the microwave network given in Fig. 39 can be written as [30]

$$\begin{aligned} S_{lm} &= \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} = \begin{bmatrix} 0 & \exp(-\gamma_m l_m) \\ \exp(-\gamma_m l_m) & 0 \end{bmatrix} \\ S_1 &= \frac{1}{\gamma_a + \gamma_1} \begin{bmatrix} \gamma_0 - \gamma_1 & 2\sqrt{\gamma_a\gamma_1} \\ 2\sqrt{\gamma_a\gamma_1} & \gamma_0 - \gamma_1 \end{bmatrix} \\ S_2 &= \frac{1}{\gamma_2 + \gamma_a} \begin{bmatrix} \gamma_2 - \gamma_a & 2\sqrt{\gamma_a\gamma_2} \\ 2\sqrt{\gamma_a\gamma_2} & \gamma_2 - \gamma_a \end{bmatrix} \end{aligned} \tag{91}$$

where l_m is the length of each homogeneous section. γ_m is the propagation constant in different sections of the network and is given as

$$\gamma_m = \begin{cases} j\sqrt{\omega^2\mu_m\varepsilon_m - \left(\dfrac{\pi}{a}\right)^2} & \text{if } \omega^2\mu_m\varepsilon_m \geq \left(\dfrac{\pi}{a}\right)^2 \\ \sqrt{\left(\dfrac{\pi}{a}\right)^2 - \omega^2\mu_m\varepsilon_m} & \text{otherwise} \end{cases} \tag{92}$$

Using (84) the T matrix of the microwave network can be written as

$$T_{lm} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} = \begin{bmatrix} \exp(-\gamma_m l_m) & 0 \\ 0 & \exp(+\gamma_m l_m) \end{bmatrix}$$

$$T_1 = \frac{1}{2\sqrt{\gamma_a \gamma_1}} \begin{bmatrix} \gamma_a + \gamma_1 & \gamma_a - \gamma_1 \\ \gamma_a - \gamma_1 & \gamma_a + \gamma_1 \end{bmatrix} \tag{93}$$

$$T_2 = \frac{1}{2\sqrt{\gamma_a \gamma_2}} \begin{bmatrix} \gamma_2 + \gamma_a & \gamma_2 - \gamma_a \\ \gamma_2 - \gamma_a & \gamma_2 + \gamma_a \end{bmatrix}$$

The overall T matrix of the two-port network using (83) is given by

$$\begin{aligned} T_{11} &= \frac{1}{4\gamma_a\sqrt{\gamma_2\gamma_1}}[(\gamma_2+\gamma_a)(\gamma_a-\gamma_1)\exp(-\gamma_a l_a) + (\gamma_2-\gamma_a) \\ &\quad \times (\gamma_a-\gamma_1)\exp(+\gamma_a l_a)]\exp(-\gamma_1 l_1)\exp(\gamma_2 l_2) \\ T_{12} &= \frac{1}{4\gamma_a\sqrt{\gamma_2\gamma_1}}[(\gamma_2+\gamma_a)(\gamma_a-\gamma_1)\exp(-\gamma_a l_a) + (\gamma_2-\gamma_a) \\ &\quad \times (\gamma_a+\gamma_1)\exp(+\gamma_a l_a)]\exp(+\gamma_1 l_1)\exp(-\gamma_2 l_2) \\ T_{21} &= \frac{1}{4\gamma_a\sqrt{\gamma_2\gamma_1}}[(\gamma_2-\gamma_a)(\gamma_a+\gamma_1)\exp(-\gamma_a l_a) + (\gamma_2+\gamma_a) \\ &\quad \times (\gamma_a-\gamma_1)\exp(+\gamma_a l_a)]\exp(-\gamma_1 l_1)\exp(-\gamma_2 l_2) \\ T_{22} &= \frac{1}{4\gamma_a\sqrt{\gamma_2\gamma_1}}[(\gamma_2-\gamma_a)(\gamma_a-\gamma_1)\exp(-\gamma_a l_a) + (\gamma_2+\gamma_a) \\ &\quad \times (\gamma_a+\gamma_1)\exp(+\gamma_a l_a)]\exp(+\gamma_1 l_1)\exp(\gamma_2 l_2) \end{aligned} \tag{94}$$

Substituting (90) in (86) leads to the expression for S_{22}

$$\begin{aligned} S_{22}^0 = -\frac{T_{21}^0}{T_{22}^0} &= \frac{\begin{matrix}[(\gamma_2+\gamma_a)(\gamma_a-\gamma_1)\exp(-\gamma_a l_a) \\ +(\gamma_2-\gamma_a)(\gamma_a+\gamma_1)\exp(+\gamma_a l_a)]\end{matrix}}{\begin{matrix}[(\gamma_2-\gamma_a)(\gamma_a-\gamma_1)\exp(-\gamma_a l_a) \\ +(\gamma_2+\gamma_a)(\gamma_a+\gamma_1)\exp(+\gamma_a l_a)]\end{matrix}} \exp(-2\gamma_1 l_1) \\ &= \Gamma_a^* \end{aligned} \tag{95}$$

The length of the airgap is used to tune the necessary condition given by (82). Figures 40 and 41 show the amplitude and phase of S_{22} as a function of the airgap length and with l_1 as parameter. The frequency is 1.8 GHz. Note that the length of l_1 does not have any effect on the amplitude of S_{22}. The dielectric material has a dielectric constant of 2.53. The flowchart in Fig. 42 gives the procedure for matching the aperture reflection. Using (81) and (84), the input reflection is related to the T matrix as follows:

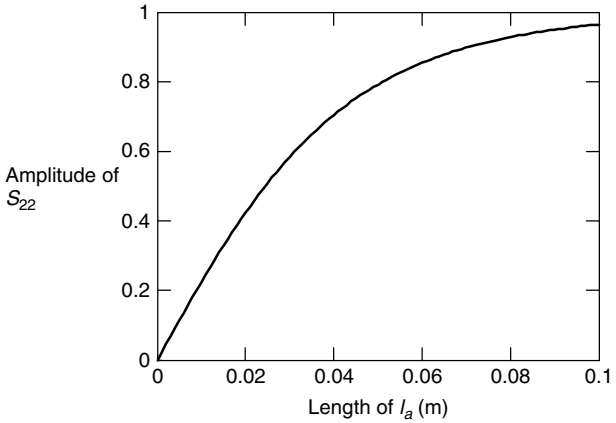


Figure 40. The amplitude of S_{22} as a function of l_a.

$$\Gamma_i = \frac{b_1}{a_1} = \frac{T_{11}^0 \Gamma_a + T_{12}^0}{T_{21}^0 \Gamma_a + T_{22}^0} \tag{96}$$

where the elements of the T matrix are as given by Eq. (90). Figure 43 shows the input reflection as a function of frequency in the L and X bands. The dielectric has a constant of 2.53. The length of the airgap is a parameter.

The length of l_1 is used to tune the minimum of the input reflection for a desired frequency. Figure 44 shows the input reflection as function of frequency for different values of l_1. From this figure the designer can choose the length, l_1, for tuning the resonance frequency.

6. *E*-PLANE STEPPED DFW

In many microwave applications it is necessary to have waveguides with larger aperture dimensions in order to

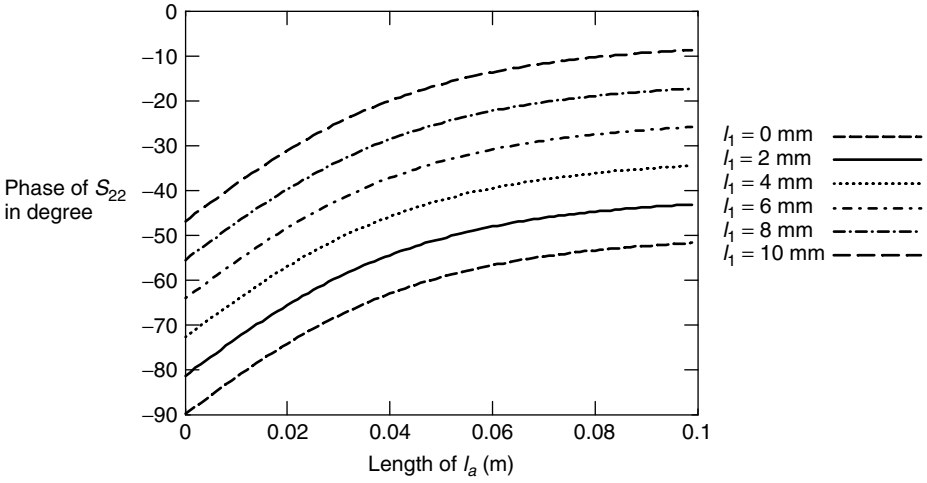


Figure 41. The phase of S_{22} as function of l_a with l_1 as parameter.

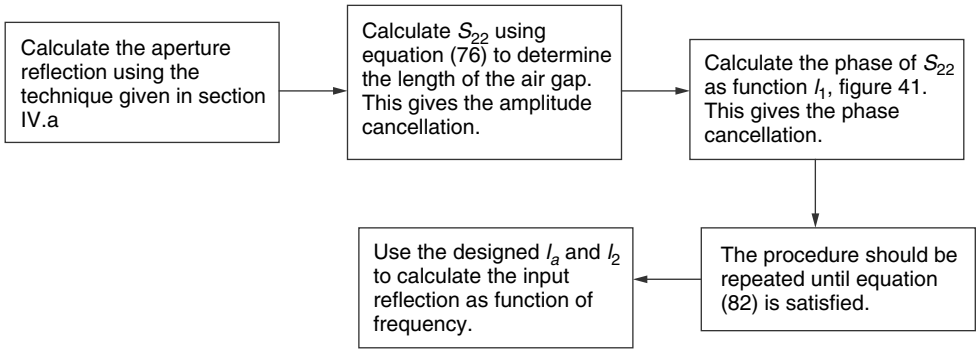


Figure 42. Flowchart for designing an airgap matching network.

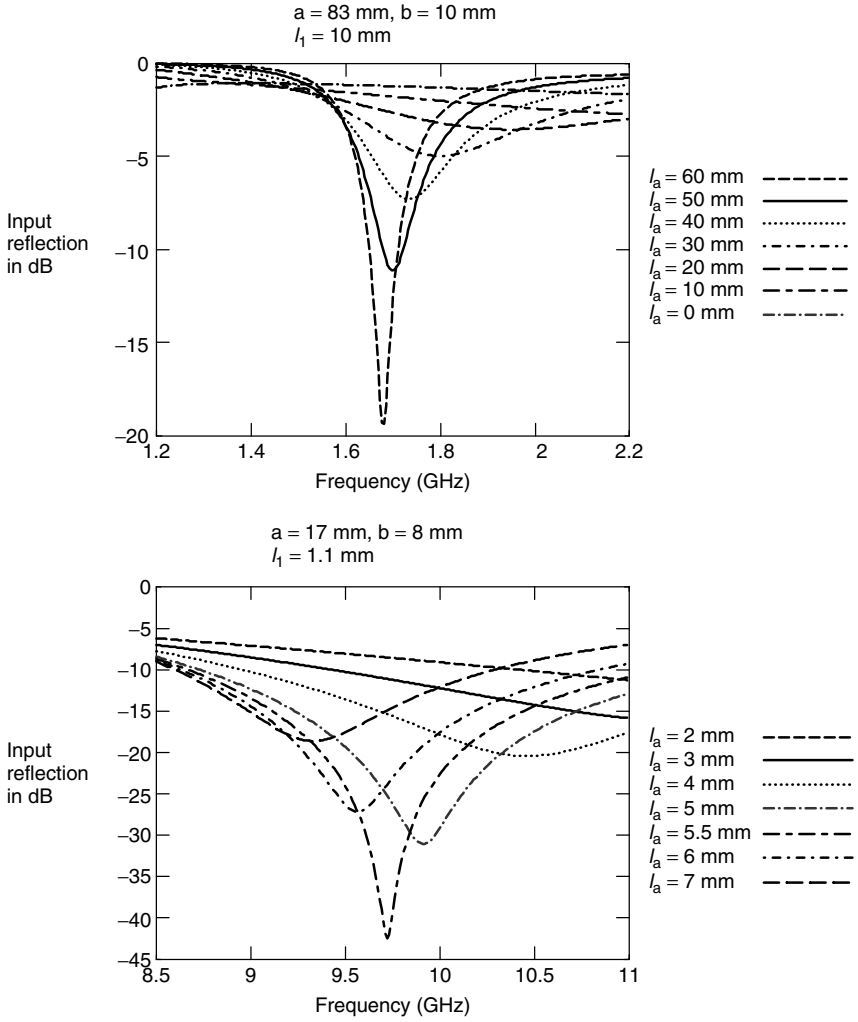


Figure 43. The input reflection as function of frequency in the L and X bands. The length of the airgap is the parameter.

have a better gain and to be able to integrate the antenna with a planar circuit. To achieve this goal, waveguide step discontinuities have been suggested [23]. The electromagnetic boundary conditions at a discontinuity usually require the presence of high-order modes. When the new dimensions of the waveguide are such that the higher-order modes are below cutoff, these modes are confined to a region very close to the discontinuity. A reactive network can then model these localized modes. Figure 45 illustrates the steps in height (*E*-plane stepped) and width (*H*-plane stepped).

The discontinuities in the waveguide can be used to realize matching networks, phase shifters, high- or lowpass filters, and resonators. In this section the *S-T* matrix approach and the airgap matching techniques discussed in Section 5 are used to characterize the input

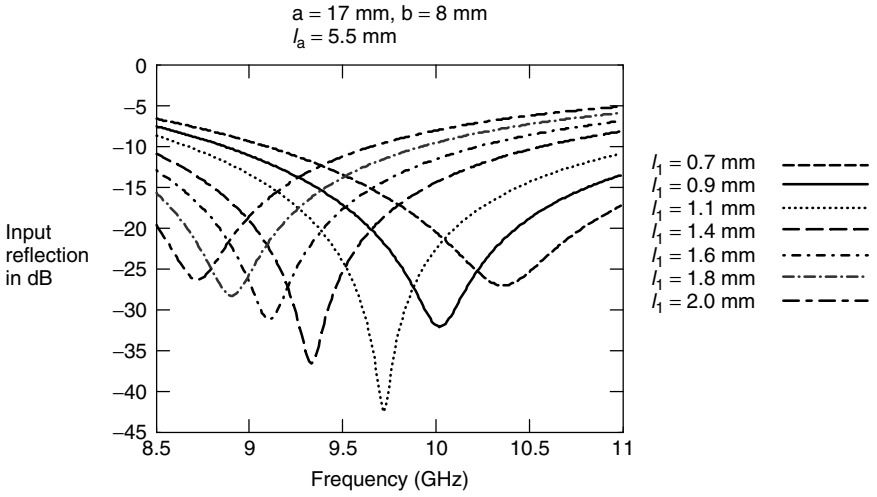


Figure 44. The input reflection as a function of frequency in the X band. The length l_1 is the parameter.

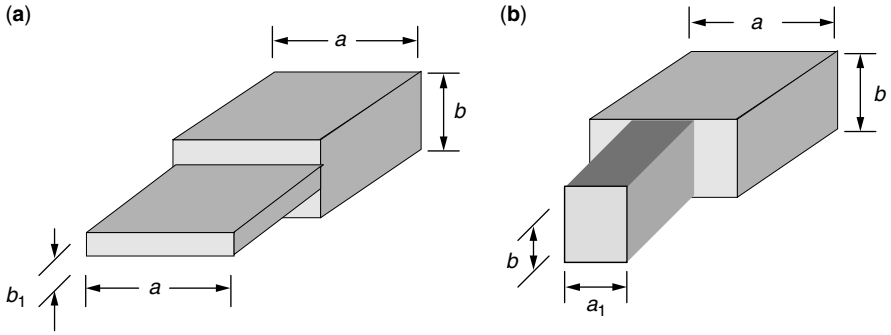


Figure 45. Symmetric step discontinuities in a waveguide: (**a**) E plane; (**b**) H plane.

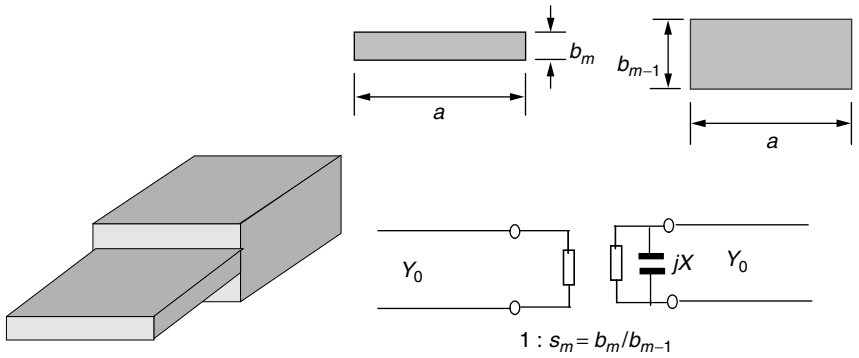


Figure 46. The stepped E plane and its lumped-circuit network representation.

reflection of a double-stepped E-plane DFW. The E-plane step configuration, its dimensions, and the lumped-circuit network are shown in Fig. 46.

The lumped-circuit network representation is a capacitance, and its susceptance value is given by [31]

$$X_m = Y_{10}\frac{2b_m}{\lambda_g}\left\{\ln\left(\frac{1-s_{m+1}^2}{4s_{m+1}}\right)\left(\frac{1+s_{m+1}}{1-s_{m+1}}\right)^{1/2(s_{m+1}+(1/s_{m+1}))} + \frac{2}{H_m}\right\}$$

$$\lambda_g = \frac{2\pi}{\gamma_m} = \frac{2\pi}{\sqrt{\omega^2\mu_m\varepsilon_m - \left(\frac{\pi}{a}\right)^2}}$$

$$s_m = \frac{b_m}{b_{m-1}}, H_m = \left(\frac{1+s_{m+1}}{1-s_{m+1}}\right)^{2s_{m+1}} \times \frac{1+\sqrt{1-\left(\frac{\gamma_m b_m}{2\pi}\right)^2}}{1-\sqrt{1-\left(\frac{\gamma_m b_m}{2\pi}\right)^2}} - \frac{1-3s_{m+1}^2}{1-s_{m+1}^2}$$

$$Y_{10} = \frac{\gamma_m}{\omega\mu} \quad m = 1, 2 \tag{97}$$

where Y_{10} is the waveguide admittance of the TE_{10} mode. b_{m-1} and b_m are the heights (steps) of the waveguides, λ_g is the wavelength of the dominant mode in the waveguide,

and S_m is the step ratio. It is shown that an accurate result can be achieved by considering only the fundamental mode. The transmission matrix elements of the E-plane step are

$$T_{11}^m = \frac{1 + S_{m+1} - jX_m \dfrac{\gamma_m b_m}{\pi} S_{m+1}}{2\sqrt{S_{m+1}}},$$

$$T_{12}^m = \frac{1 - S_{m+1} - jX_m \dfrac{\gamma_m b_m}{\pi} S_{m+1}}{2\sqrt{S_{m+1}}}$$

$$T_{21}^m = \frac{1 - S_{m+1} + jX_m \dfrac{\gamma_m b_m}{\pi} S_{m+1}}{2\sqrt{S_{m+1}}},$$

$$T_{22}^m = \frac{1 + S_{m+1} + jX_m \dfrac{\gamma_m b_m}{\pi} S_{m+1}}{2\sqrt{S_{m+1}}} \quad m = 1, 2 \tag{98}$$

where $m = 1,2$ for the first and second steps, respectively. The input reflection at the reference plane (see Fig. 47) is given by Eq. (92), where the total T matrix is given by

$$T = T_{l4}T_4T_{l3}T_3T_{l2}T_2T_{la}T_1T_{l1} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \tag{99}$$

Using Eqs. (95), (88), (89), and (94), the elements of the overall transmission matrix are

$$T_{11} = \frac{1}{4\gamma_0\sqrt{\gamma_1\gamma_2}} \left\{ \begin{array}{l} \{[A_1T_{11}^1A_2T_{11}^2 \\ \quad + A_1T_{12}^1B_2T_{21}^2]A_3C \\ \quad + [A_1T_{11}^1A_2T_{12}^2 \\ \quad + A_1T_{12}^1B_2T_{22}^2]B_3E\}A_4G \\ \quad + \{[A_1T_{11}^1A_2T_{11}^2 \\ \quad + A_1T_{12}^1B_2T_{21}^2]A_3D \\ \quad + [A_1T_{11}^1A_2T_{12}^2 \\ \quad + A_1T_{12}^1B_2T_{22}^2]B_3F\}B_4I \end{array} \right\} A_5$$

$$T_{12} = \frac{1}{4\gamma_0\sqrt{\gamma_1\gamma_2}} \left\{ \begin{array}{l} \{[A_1T_{11}^1A_2T_{11}^2 \\ \quad + A_1T_{12}^1B_2T_{21}^2]A_3C \\ \quad + [A_1T_{11}^1A_2T_{12}^2 \\ \quad + A_1T_{12}^1B_2T_{22}^2]B_3E\}A_4H \\ \quad + \{[A_1T_{11}^1A_2T_{11}^2 \\ \quad + A_1T_{12}^1B_2T_{21}^2]A_3D \\ \quad + [A_1T_{11}^1A_2T_{12}^2 \\ \quad + A_1T_{12}^1B_2T_{22}^2]B_3F\}B_4J \end{array} \right\} B_5$$

$$T_{21} = \frac{1}{4\gamma_0\sqrt{\gamma_1\gamma_2}} \left\{ \begin{array}{l} \{[B_1T_{21}^1A_2T_{11}^2 \\ \quad + B_1T_{22}^1B_2T_{21}^2]A_3C \\ \quad + [B_1T_{21}^1A_2T_{12}^2 \\ \quad + B_1T_{22}^1B_2T_{22}^2]B_3E\}A_4G \\ \quad + \{[B_1T_{21}^1A_2T_{11}^2 \\ \quad + B_1T_{22}^1B_2T_{21}^2]A_3D \\ \quad + [B_1T_{21}^1A_2T_{12}^2 \\ \quad + B_1T_{22}^1B_2T_{22}^2]B_3F\}B_4I \end{array} \right\} A_5$$

$$T_{22} = \frac{1}{4\gamma_0\sqrt{\gamma_1\gamma_2}} \left\{ \begin{array}{l} \{[B_1T_{21}^1A_2T_{11}^2 \\ \quad + B_1T_{22}^1B_2T_{21}^2]A_3C \\ \quad + [B_1T_{21}^1A_2T_{12}^2 \\ \quad + B_1T_{22}^1B_2T_{22}^2]B_3E\}A_4H \\ \quad + \{[B_1T_{21}^1A_2T_{11}^2 \\ \quad + B_1T_{22}^1B_2T_{21}^2]A_3D \\ \quad + [B_1T_{21}^1A_2T_{12}^2 \\ \quad + B_1T_{22}^1B_2T_{22}^2]B_3F\}B_4J \end{array} \right\} B_5 \tag{100}$$

where

$$T_{11}^1 = \frac{1 + S_2 - jB_1 \dfrac{\gamma_2 b_1}{\pi} S_2}{2\sqrt{S_2}}, \quad T_{12}^1 = \frac{1 - S_2 - jB_1 \dfrac{\gamma_2 b_1}{\pi} S_2}{2\sqrt{S_2}},$$

$$T_{21}^1 = \frac{1 - S_2 + jB_1 \dfrac{\gamma_2 b_1}{\pi} S_2}{2\sqrt{S_2}}, \quad T_{22}^1 = \frac{1 + S_2 + jB_1 \dfrac{\gamma_2 b_1}{\pi} S_2}{2\sqrt{S_2}}$$

$$T_{11}^2 = \frac{1 + S_3 - jB_2 \dfrac{\gamma_3 b_2}{\pi} S_3}{2\sqrt{S_3}}, \quad T_{12}^2 = \frac{1 - S_3 - jB_2 \dfrac{\gamma_3 b_2}{\pi} S_3}{2\sqrt{S_3}},$$

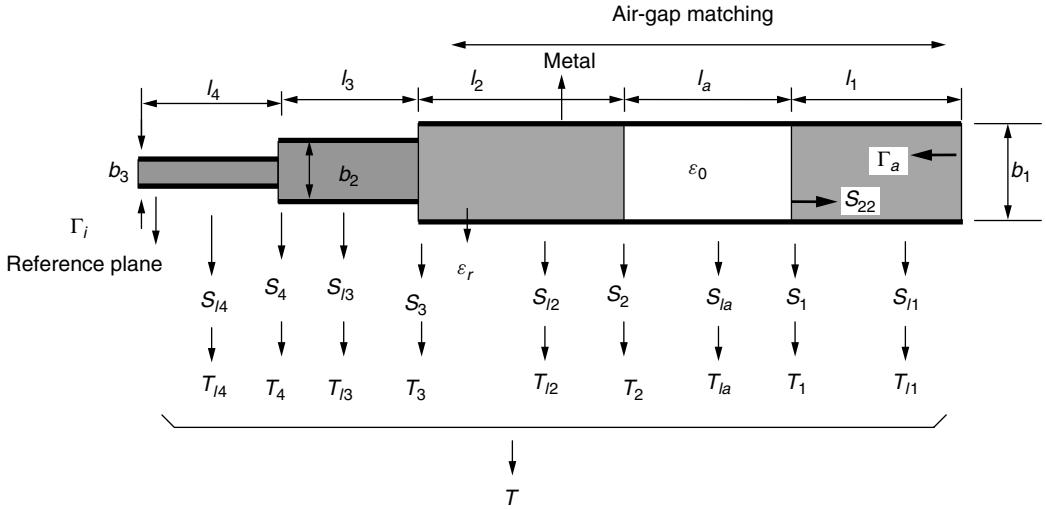


Figure 47. DFW with the airgap matching network and two E plane steps as a cascaded S–T network.

$$T_{12}^2 = \frac{1 - S_3 + jB_2\frac{\gamma_3 b_2}{\pi}S_3}{2\sqrt{S_3}}, \quad T_{22}^2 = \frac{1 + S_3 + jB_2\frac{\gamma_3 b_2}{\pi}S_3}{2\sqrt{S_3}}$$

$$A_1A_2 = \exp(-\gamma_4 l_4)\exp(-\gamma_3 l_3),$$

$$A_1B_2 = \exp(-\gamma_4 l_4)\exp(+\gamma_3 l_3)$$

$$A_3C = \exp(-\gamma_2 l_2)(\gamma_2 + \gamma_a), \quad B_3E = \exp(+\gamma_2 l_2)(\gamma_2 - \gamma_a),$$

$$A_4G = \exp(-\gamma_a l_a)(\gamma_a + \gamma_1) \quad A_3D = \exp(-\gamma_2 l_2)(\gamma_2 - \gamma_a),$$

$$B_3F = \exp(+\gamma_2 l_2)(\gamma_2 + \gamma_a), \quad B_4I = \exp(+\gamma_a l_a)(\gamma_a - \gamma_1)$$

$$A_4H = \exp(-\gamma_a l_a)(\gamma_a - \gamma_1), \quad B_4J = \exp(+\gamma_a l_a)(\gamma_a + \gamma_1)$$

$$B_1A_2 = \exp(+\gamma_4 l_4)\exp(-\gamma_3 l_3),$$

$$B_1B_2 = \exp(+\gamma_4 l_4)\exp(+\gamma_3 l_3),$$

$$A_5 = \exp(-\gamma_1 l_1), B_5 = \exp(+\gamma_1 l_1) \tag{101}$$

where $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4$, since the dielectric constant of the filling material is the same for different sections. The input reflection can be calculated as a function of different parameters. These parameters are the frequency, the dielectric constant, the length of the airgap, and the length of the filled homogeneous section. Figure 48 shows the input reflection as a function of frequency for different parameters at reference plane 1 (see Fig. 49).

For this case $a_1 = a_2 = a_3 = 17$ mm, $b_1 = 11$ mm, $b_2 = 8$ mm, $b_3 = 5$ mm, and $\varepsilon_r = 2.53$. Note that the aperture reflection is calculated using the method given in Section 4. The length of l_4 does not affect the amplitude of the input reflection coefficient. These optimal values are calculated by the trial-and-error method. Note that the input reflection is very sensitive to the length of different sections.

6.1. Measurement Results

Figure 49 shows the realized E-plane stepped DFW in the X band, $a_1 = a_2 = a_3 = 17$ mm and $\varepsilon_r = 2.53$.

In order to carry out the input reflection measurement at the reference plane 1 (see Fig. 49), the antenna is calibrated using a modified waveguide calibration technique [32]. Three standard short circuits with different offset delays are designed for this purpose. Figure 50 shows the calibration set. The width of the three waveguide standards equals $a = 17$ mm.

Figure 51 compares the calculated and measured input reflections at the reference plane 1 as a function of frequency.

The measurement differs from the theoretical results since the two steps were not constructed in one piece. In order to ensure good galvanic contact between the waveguide steps, it was necessary to use screws. Such a technique can lead to an airgap between the steps. Also the inaccuracies in length of each section were not

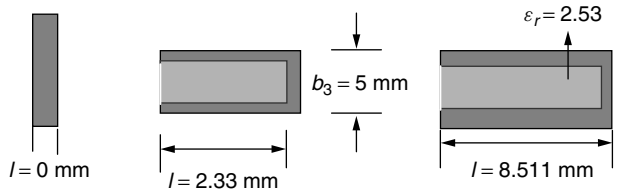


Figure 50. The standard short circuits for waveguide calibration.

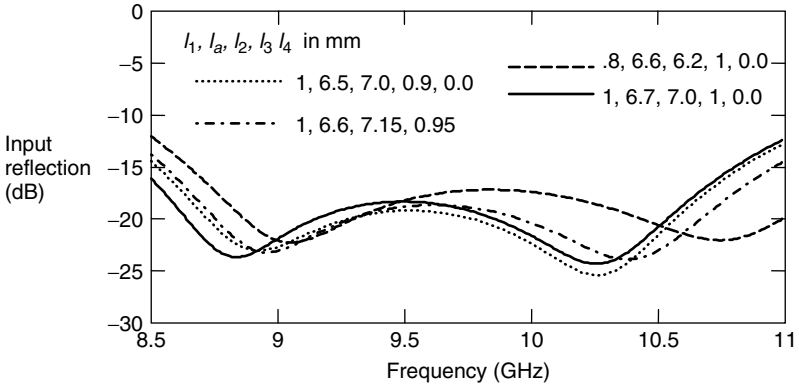


Figure 48. The input reflection as a function of frequency.

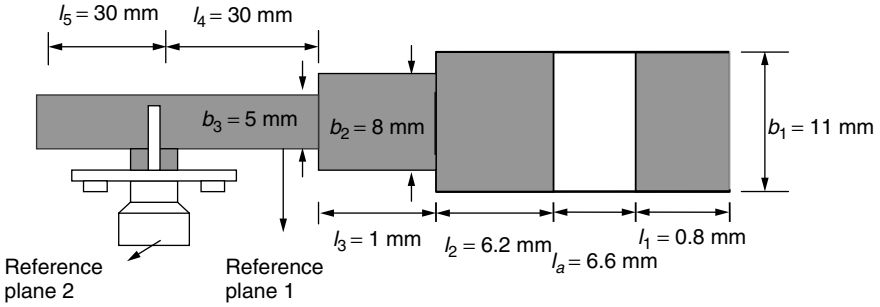


Figure 49. Realized DFW in X band.

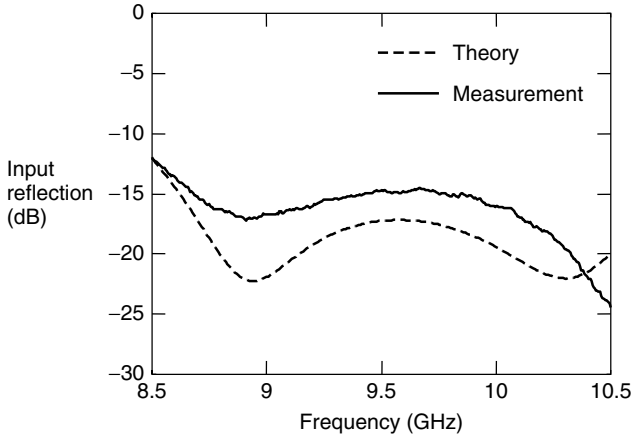


Figure 51. Comparison between the calculated and measured input reflection at reference plane 1.

taken into account. It can be seen from Fig. 48 that the input reflection is very sensitive to the values of the design parameters.

7. DUAL-POLARIZED WAVEGUIDE ANTENNA

In many applications polarization plays a major role. Since the late 1990s there has been an enormous expansion in wireless communication systems and in the number of people using their services. Limited bandwidth is the major concern that may prevent expansion of the capacity. In order to expand the capacity, the use of multiple antennas at the base station has been suggested [33]. Such antennas consist of a number of antenna elements that can transmit and receive the signals independently from each other. Using signal processing algorithms, multiple antenna systems can continuously distinguish between the desired signals and multipath and interfering signals by tracking the desired users with the main lobes and the interferers with the pattern minima. In this way it is possible to maximize the carrier-to-interference ratio (C/I) and the signal-to-noise-ratio (SNR). This is called *space-division multiplexing access* (SDMA). If dual-polarized antennas are used so that two orthogonal polarizations are received, this can enhance the array signal processing in two different ways. First, the number of available signals is increased, which can further improve C/I after signal processing [34,35]. It is also possible to use a postdetection maximum ratio combining technique based on signals coming from the two polarizations.

Radar polarimetry is a valuable technique for the extraction of geophysical parameters from synthetic aperture radar (SAR) images and terrain classification. In many radar applications SAR is used for target detection, classification, and identification. Cloude and Papathanassiou [36] describe the use of a dual-polarized antenna system in a spaceborne satellite in NASA's mission to the planet earth that is intended to provide measurements of the earth's environment (air, water, land). The measurements are used to determine land–surface soil moisture, ocean salinity, surface temperature, and vegetation water content in the L band (1.4 GHz). Liu et al. [37] use a high-resolution dual-polarization X-band radar at "low grazing" angle to obtain images from the ocean surface. Characteristics of low-grazing-angle backscatter marked differences in horizontally and vertically polarized Doppler properties.

7.1. Design

When designing dual polarized antennas, one needs to keep the input reflection of both polarizations at the coax reference plane and the aperture reflection as low as possible. At the same time the isolation between the two feeds must be as high as possible. The techniques described in the previous sections can be used to match the aperture reflection and the coax-to-waveguide transitions. For the isolation between the two polarizations, a polarization filter needs to be designed.

It is desired for the filter to be transparent for one polarization and to reflect the other one. Since the waveguide cutoff frequency depends on polarization, it is possible to design a polarization filter that satisfies this requirement. The polarization filter is a piece of thin metal, which is located vertically or horizontally inside the waveguide. In this way the piece of metal divides the waveguide section in two new ones, each with a different cut off frequency than the original one. The concept is illustrated in Fig. 54.

The polarization filter divides the original waveguide in two equal waveguides in which only the fundamental mode can be propagated with vertical polarization. The longer the filter length, the more the unwanted polarization will be attenuated. The thinner the filter, the less it would disturb the fundamental vertical mode and thus the more transparent it becomes for the desired polarization.

In order to design the length of the filter, one must to calculate the attenuation of the mode in the waveguide after the polarization filter. Using Eq. (21), for TE_{10}, the cutoff frequency of the waveguide given in Fig. 52 becomes

$$f_{c_{10}} = \frac{c}{2a} \tag{102}$$

Substituting (98) and $f = c/\lambda$ in Eq. (22), the propagation constant for the evanescent waves (horizontal polarization) becomes

$$\begin{aligned} \beta_z &= \beta\sqrt{\left(\frac{f_c}{f}\right)^2 - 1} = \frac{2\pi}{\lambda}\sqrt{\left(\frac{\lambda}{2b}\right)^2 - 1} \\ (\beta_z)^2 &= (2\pi)^2\left[\left(\frac{1}{2b}\right)^2 - \left(\frac{1}{\lambda}\right)^2\right] \end{aligned} \tag{103}$$

Consider

$$\delta = 20\log_{10}[e^{\beta_z \rho}] \tag{104}$$

where ρ is the length of the filter and δ is the attenuation of the wave in decibels. This becomes

$$\delta = \frac{20}{\ln(10)}\ln[e^{\beta_z \rho}] = \frac{20}{\ln(10)}\beta_z \rho \tag{105}$$

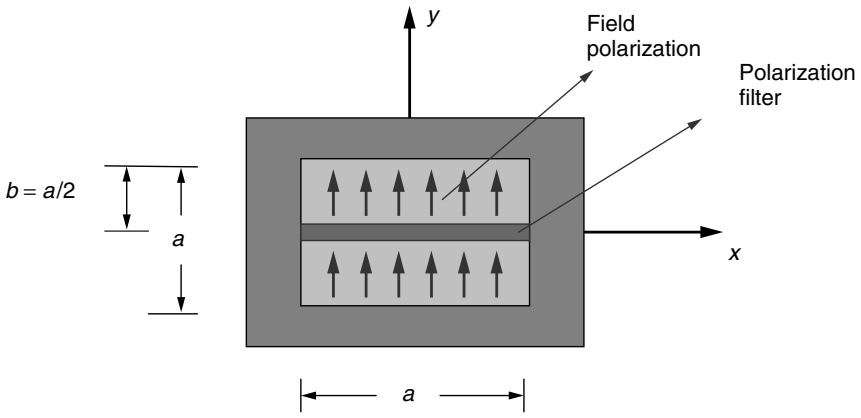


Figure 52. Front view of the polarization filter [38].

Substituting Eq. (99) into (101) leads to

$$\delta = \frac{20\rho}{\ln(10)}(2\pi)\sqrt{\left(\frac{1}{2b}\right)^2 - \left(\frac{1}{\lambda}\right)^2} \tag{106}$$

Rearranging (102) gives

$$\frac{\delta}{\rho} = \frac{40\pi}{\ln(10)}\frac{1}{\lambda}\sqrt{\left(\frac{\lambda}{2b}\right)^2 - 1} \tag{107}$$

An upper bound can be found and is given by the following equation:

$$\frac{\delta}{\rho} \le \frac{40\pi}{\ln(10)}\frac{1}{2b} \le 27.3\frac{1}{b} \tag{108}$$

Figure 53 shows the attenuation of the TE_{10} mode as a function of the length of the polarization filter for three different dielectric materials. It is shown that as the filter length increases, the field attenuates more.

Figure 54 shows the layout of a dual-polarized DFW with a polarization filter and an airgap matching network. In order to lower the diffraction effect from the edge of the waveguide, edge tapering is introduced.

Figure 55 shows the calibrated measurement results for input reflection and isolation after tuning. The *E*- and *H*-plane radiation patterns are shown in Fig. 56. The pattern computation is given in Section 8.

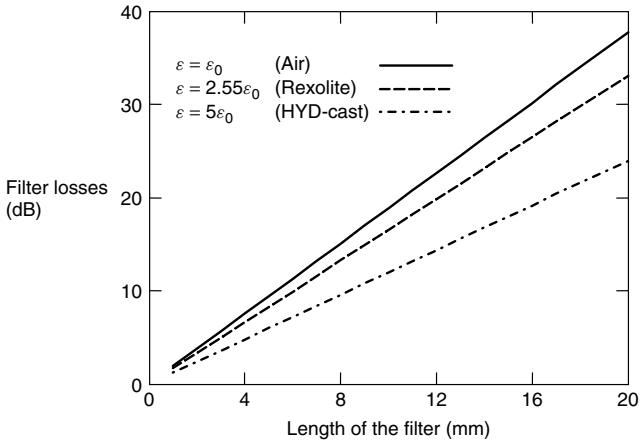


Figure 53. The attenuation of TE_{10} mode as function of polarization filter length at $f = 4$ GHz.

8. RADIATION PATTERN

The waveguide antenna is considered as an aperture antenna. The far-field radiation pattern of the waveguide can be calculated using the equivalent principle [40]. It states that based on the known electromagnetic modes propagating inside the waveguide, one can construct the electric and magnetic currents on the aperture plane. The

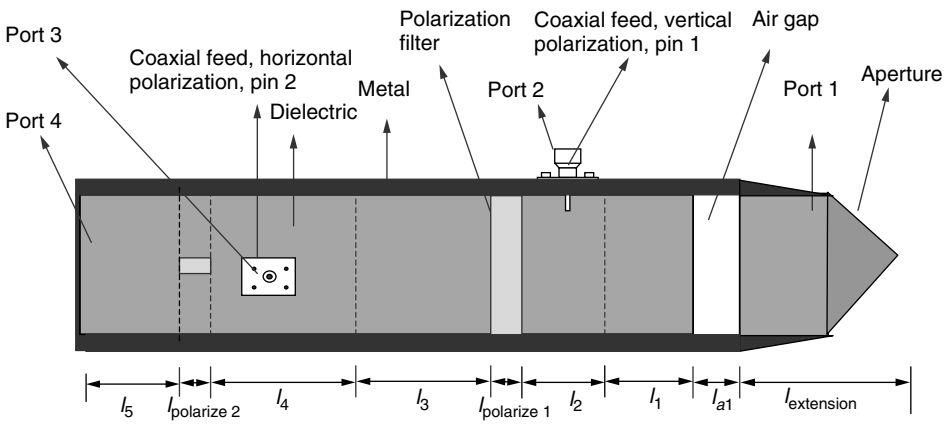


Figure 54. Dual-polarized dielectric-filled waveguide with polarization filter and edge tapering in S-band, side view. (Courtesy of IRCTR.)

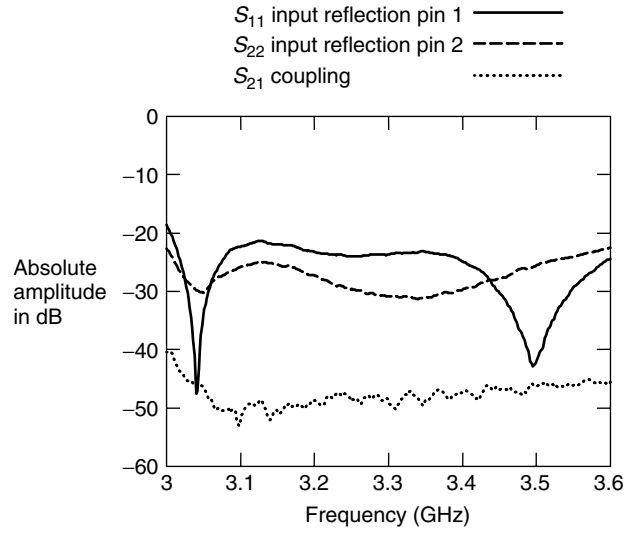


Figure 55. The measured input reflection and isolation of the dual-polarized dielectric-filled waveguide.

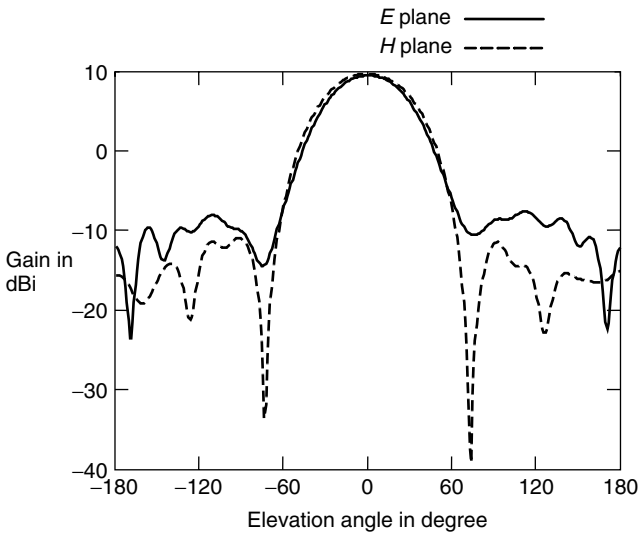


Figure 56. The measured E and H planes of the dielectric-filled waveguide at S band, $f = 3.3$ GHz.

electric and magnetic currents can then be used to set up the "potentials" integral equations to calculate the far-field radiation pattern.

The electromagnetic potentials are a solution to the vector wave equation and can be written as [40]

$$\mathbf{A} = \frac{\mu}{4\pi} \iint_s \mathbf{J} \frac{e^{-jkR}}{R}\, ds'$$
$$\mathbf{F} = \frac{\varepsilon}{4\pi} \iint_s \mathbf{M} \frac{e^{-jkR}}{R}\, ds' \tag{109}$$

where $\mathbf{J}$ and $\mathbf{M}$ are the electric and magnetic current sources; ds' is a differential area. The coordinate system to analyze the radiation pattern is shown in Fig. 57.

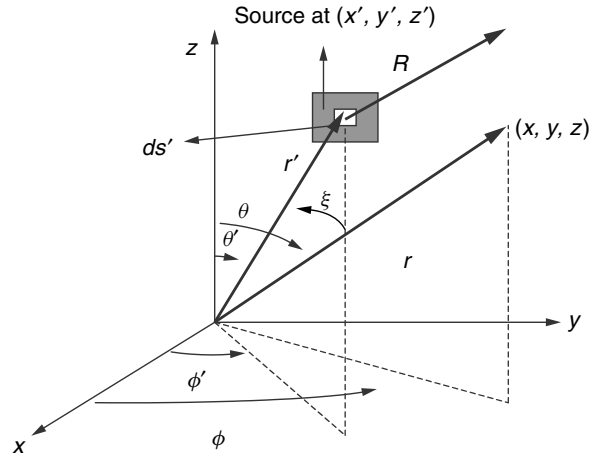


Figure 57. Coordinate system for aperture antenna analysis. (*Source:* C. Balanis.)

It is shown in [40] that in the far zone R can be approximated by

$$R \simeq r - r' \cos\xi \quad \text{for phase variations}$$
$$R \simeq r \quad \text{for amplitude variations} \tag{110}$$

where ξ is the angle between the vector r and r'. The complete electromagnetic field in the far zone is related to the electromagnetic potentials and can be written as

$$\begin{aligned} E_r &\simeq 0 \\ E_\theta &\simeq -\frac{je^{-jkr}}{4\pi r}(F_\phi + \eta A_\theta) \\ E_\phi &\simeq \frac{jke^{-jkr}}{4\pi r}(F_\theta - \eta A_\phi) \\ H_r &\simeq 0 \\ H_\theta &\simeq \frac{jke^{-jkr}}{4\pi r}\left(A_\phi - \frac{F_\theta}{\eta}\right) \\ H_\phi &\simeq -\frac{jke^{-jkr}}{4\pi r}\left(A_\theta + \frac{F_\phi}{\eta}\right) \end{aligned} \tag{111}$$

The potentials given in Eq. (107) can be derived from Eqs. (105) and (106) via

$$\begin{aligned} \mathbf{A} &= \frac{\mu e^{-jkr}}{4\pi r} \iint_s \mathbf{J} e^{jkr' \cos\xi}\, ds' \\ &= \frac{\mu e^{-jkr}}{4\pi r} \iint_s (\hat{\mathbf{a}}_x J_x + \hat{\mathbf{a}}_y J_y + \hat{\mathbf{a}}_z J_z) e^{jkr' \cos\xi}\, ds' \\ \mathbf{F} &= \frac{\varepsilon e^{-jkr}}{4\pi r} \iint_s \mathbf{M} e^{jkr' \cos\xi}\, ds' \\ &= \frac{\varepsilon e^{-jkr}}{4\pi r} \iint_s (\hat{\mathbf{a}}_x M_x + \hat{\mathbf{a}}_y M_y + \hat{\mathbf{a}}_z M_z) e^{jkr' \cos\xi}\, ds' \end{aligned} \tag{112}$$

Using the rectangular-to-spherical transformation, one can give the components of the potentials:

$$\begin{aligned}
A_\theta &= \frac{\mu e^{-jkr}}{4\pi r}\iint_s (J_x\cos\theta\cos\phi + J_y\cos\theta\sin\phi \\
&\quad - J_z\sin\theta)e^{jkr'\cos\xi}\,ds' \\
A_\phi &= \frac{\mu e^{-jkr}}{4\pi r}\iint_s (-J_x\sin\phi + J_y\cos\phi)e^{jkr'\cos\xi}\,ds' \\
F_\theta &= \frac{\varepsilon e^{-jkr}}{4\pi r}\iint_s (M_x\cos\theta\cos\phi + M_y\cos\theta\sin\phi \\
&\quad + M_z\sin\theta)e^{jkr'\cos\xi}\,ds' \\
F_\phi &= \frac{\varepsilon e^{-jkr}}{4\pi r}\iint_s (-M_x\sin\phi + M_y\cos\phi)e^{jkr'\cos\xi}\,ds'
\end{aligned} \tag{113}$$

The electric and magnetic current components in (108) and (109) in general can be found using the equivalent principle.

For a given field distribution, the analytic forms for the fields for an arrangement are not the same. However, the far-zone expression will be the same. For each geometry, the only difference in analysis is in the formulation of

1. The components of the equivalent current densities $(J_x, J_y, J_z, M_x, M_y, M_z)$
2. The difference in paths from the source to the observation point $r'\cos\xi$
3. The differential area ds'

8.1. Uniform Aperture Distribution

For simplicity, the first aperture field is considered to be constant and given by

$$E_a = \begin{cases} \hat{a}_y E_0 & -\dfrac{a}{2} \le x' \le \dfrac{a}{2}, -\dfrac{b}{2} \le y' \le \dfrac{b}{2} \\ 0 & \text{elsewhere} \end{cases} \tag{114}$$

Figure 58 shows the geometry of the waveguide aperture antenna on an infinite electric ground plane. The equivalent principle is used to form the electromagnetic current densities on the opening of the waveguide. Once the current densities are calculated, Eqs. (107) and (109) are used to determine the components of the electromagnetic field in the far zone. The components of the current densities are given by

$$M_s = \begin{cases} -2\hat{n}\times E_a = -2\hat{a}_z\times\hat{a}_y E_0 \\ \qquad\quad = +2\hat{a}_x E_0 & -\dfrac{a}{2} \le x' \le \dfrac{a}{2}, \\ & -\dfrac{b}{2} \le y' \le \dfrac{b}{2} \\ 0 & \text{elsewhere} \end{cases} \tag{115}$$

$$J_s = 0 \qquad \text{everywhere}$$

Substituting (111) in (109) leads to

$$\begin{aligned}
A_\theta &= A_\phi = 0 \\
F_\theta &= \frac{\varepsilon e^{-jkr}}{4\pi r}\cos\theta\cos\phi\left[\int_{-\frac{a}{2}}^{\frac{a}{2}}\int_{-\frac{b}{2}}^{\frac{b}{2}} M_x e^{jk(x'\sin\theta\cos\phi + y'\sin\theta\cos\phi)}\right. \\
&\quad \left.\times\, dx'\,dy'\right]
\end{aligned} \tag{116}$$

The integral within the brackets represents the *space factor* for a two-dimensional distribution. Using the following integral identity

$$\int_{-(c/2)}^{c/2} e^{j\alpha z}\,dz = c\left[\frac{\sin\left(\frac{\alpha}{2}c\right)}{\frac{\alpha}{2}c}\right] \tag{117}$$

Eq. (112) reduces to

$$F_\theta = \frac{\varepsilon e^{-jkr}}{4\pi r}2ab\left[\cos\theta\cos\phi\left(\frac{\sin X}{X}\right)\left(\frac{\sin Y}{Y}\right)\right] \tag{118}$$

where

$$\begin{aligned}
X &= \frac{ka}{2}\sin\theta\cos\phi \\
Y &= \frac{kb}{2}\sin\theta\sin\phi
\end{aligned} \tag{119}$$

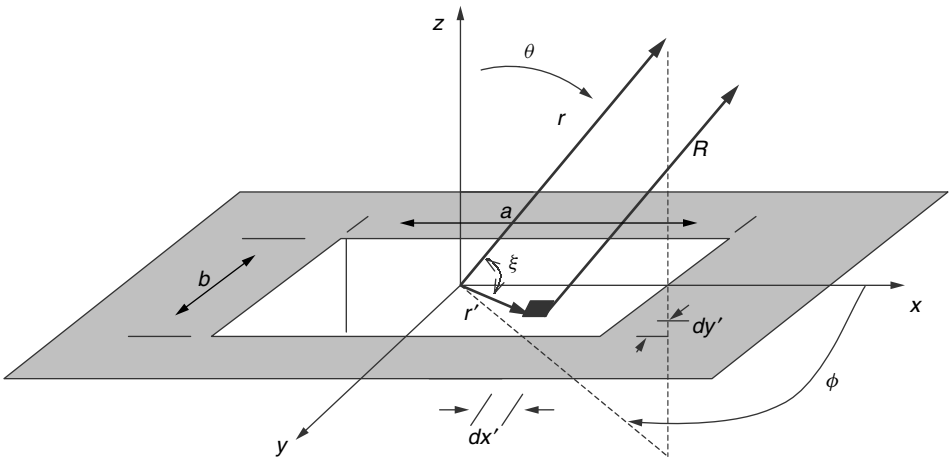


Figure 58. Rectangular waveguide aperture on an infinite electric ground plane.

Similarly it can be shown that

$$F_\phi = -\frac{\varepsilon e^{-jkr}}{4\pi r} 2ab \left[\sin\phi \left(\frac{\sin X}{X}\right)\left(\frac{\sin Y}{Y}\right)\right] \tag{120}$$

Inserting (116) and (114) in (107), the fields radiated by the waveguide aperture with uniform field distribution can be written as

$$\begin{aligned} E_r &= 0 \\ E_\theta &= -j\frac{abkE_0 e^{-jkr}}{2\pi r}\left[\sin\phi\left(\frac{\sin X}{X}\right)\left(\frac{\sin Y}{Y}\right)\right] \\ E_\phi &= j\frac{abkE_0 e^{-jkr}}{2\pi r}\left[\cos\theta\cos\phi\left(\frac{\sin X}{X}\right)\left(\frac{\sin Y}{Y}\right)\right] \\ H_r &= 0 \\ H_\theta &= -\frac{E_\phi}{\eta} \\ H_\phi &= \frac{E_\theta}{\eta} \end{aligned} \tag{121}$$

For the aperture given in Fig. 59, the E-plane pattern is on the y–z plane ($\phi = \pi/2$) and the H-plane pattern is on the x–z plane ($\phi = 0$). Thus

E plane($\phi = \pi/2$):

$$E_r = E_\phi = 0$$

$$E_\theta = j\frac{abkE_0 e^{-jkr}}{2\pi r}\left[\frac{\sin\left(\frac{kb}{2}\sin\theta\right)}{\frac{kb}{2}\sin\theta}\right] \tag{121}$$

H plane($\phi = 0$):

$$E_r = E_\theta = 0$$

$$E_\phi = j\frac{abkE_0 e^{-jkr}}{2\pi r}\left[\cos\theta\frac{\sin\left(\frac{ka}{2}\sin\theta\right)}{\frac{ka}{2}\sin\theta}\right] \tag{122}$$

Figures 59 and 60 show the three-dimensional patterns of a rectangular waveguide mounted on an infinite ground plane. Since the dimensions are greater than the wavelength, multiple lobes appear. The number of lobes is directly related to the dimension of the waveguide aperture. The pattern in the H plane is only a function of the dimension a, whereas that in the E plane is influenced only by b. In the E plane, the sidelobe formed on each side of the major lobe is a result of $\lambda < b \leq 2\lambda$. In the H plane, the first minor lobe on each side of the major lobe is formed when $\lambda < a \leq 2\lambda$ and the second sidelobe when $2\lambda < a \leq 3\lambda$. Additional lobes are formed when both aperture dimensions increase.

The patterns computed above assumed that the aperture was mounted on an infinite ground plane. In practice, infinite ground planes are not realizable. Edge effects on the patterns of apertures mounted on finite-size ground planes can be accounted for by the method of moment technique [39]. Figure 61 illustrates the electric and magnetic components of a DFW using the MoM [40].

The electric and magnetic currents are used to set up the far-field integral equations. Figure 62 shows the comparison between the simulations and measurements results for the E- and H-plane patterns.

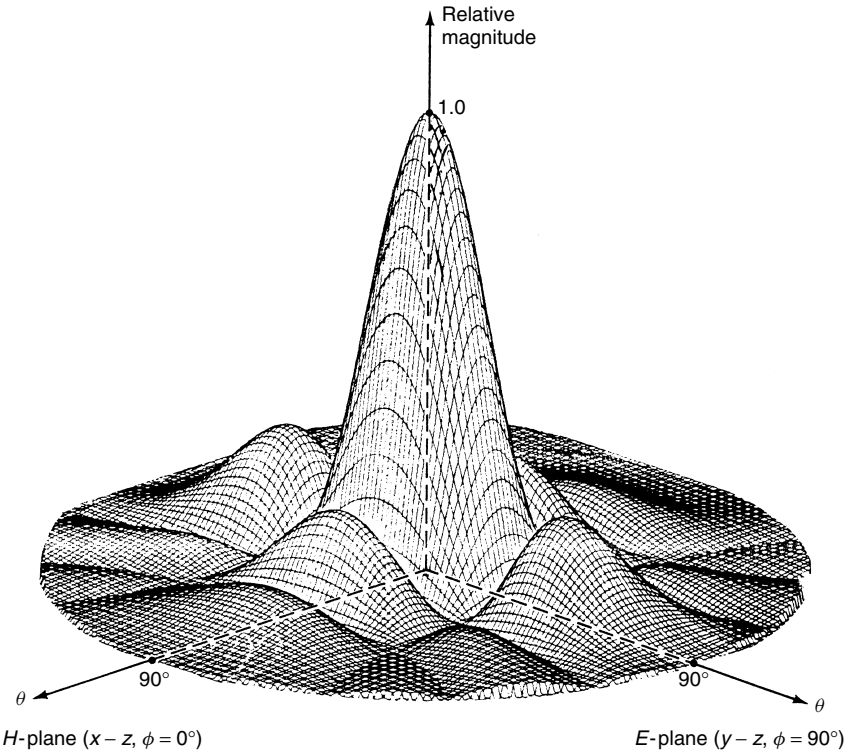


Figure 59. Three-dimensional field pattern of a constant field rectangular aperture mounted on an infinite ground plane ($a = 3\lambda$, $b = 2\lambda$).

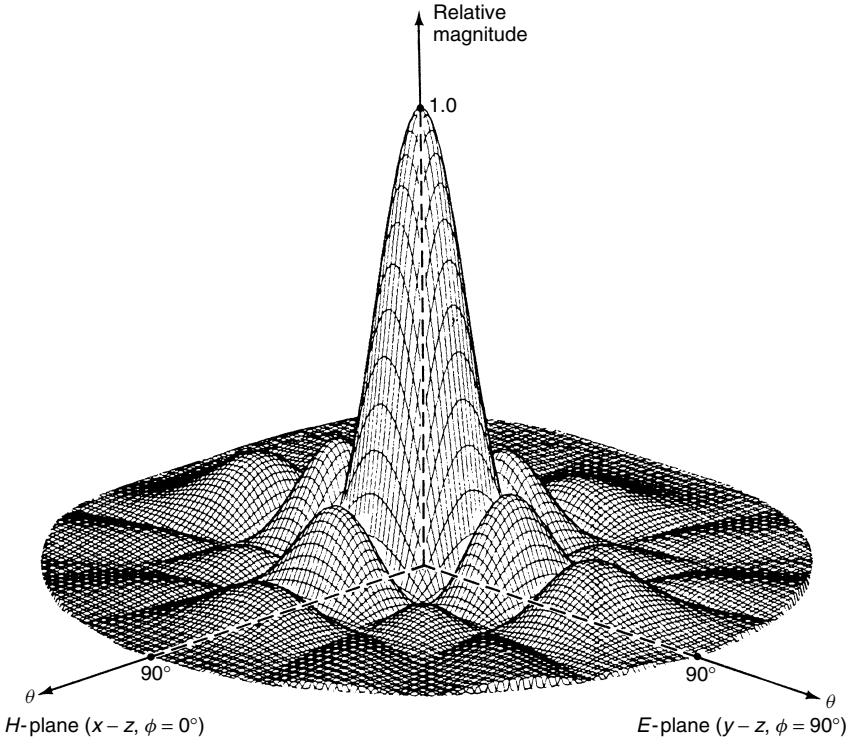


Figure 60. Three-dimensional field pattern of a constant field square aperture mounted on an infinite ground plane ($a = b = 3\lambda$).

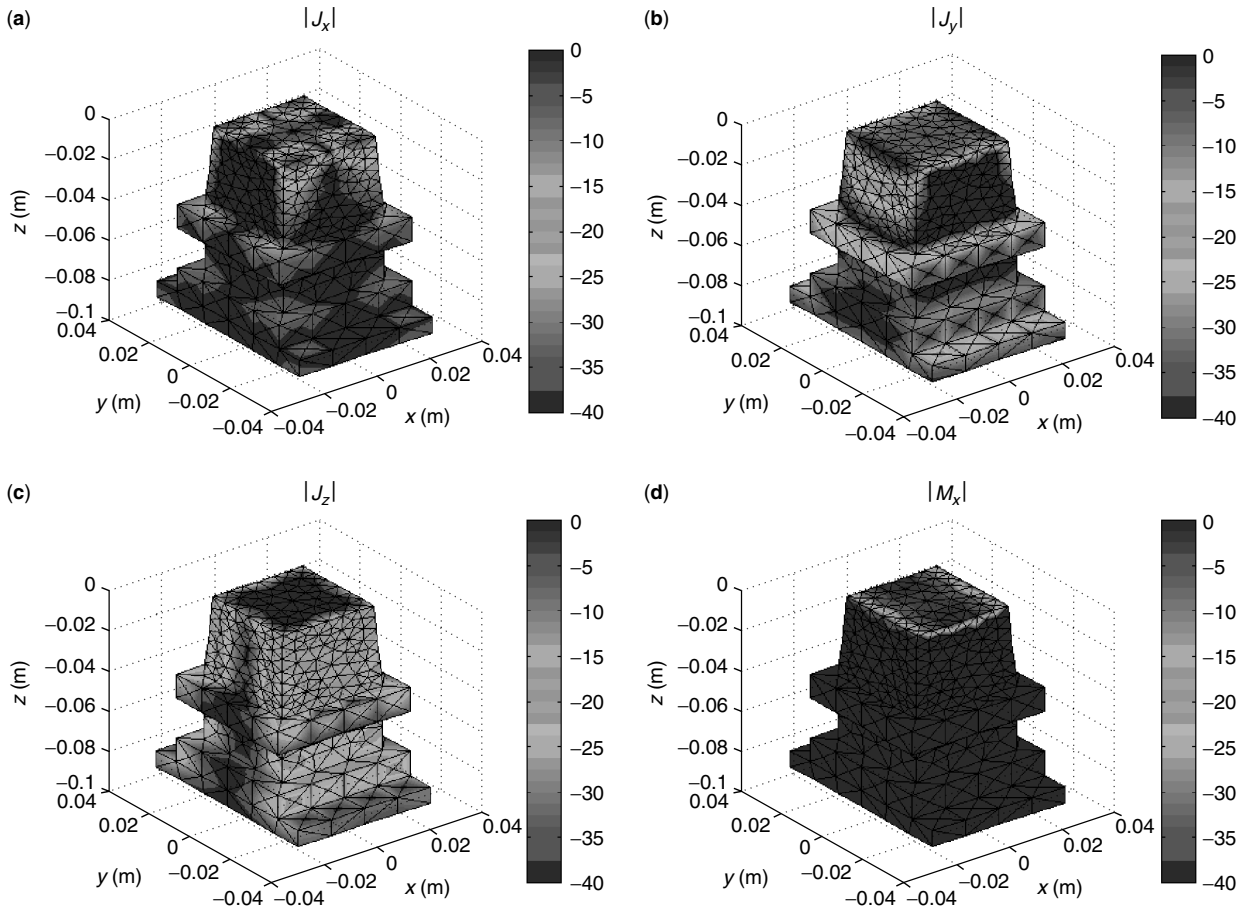


Figure 61. Relative amplitude of the different components of the induced surface currents for DFW. The electric (**a–c**) and magnetic (**d**) currents are given in decibels. Aperture dimensions: $0.38\lambda \times 0.38\lambda$ [40].

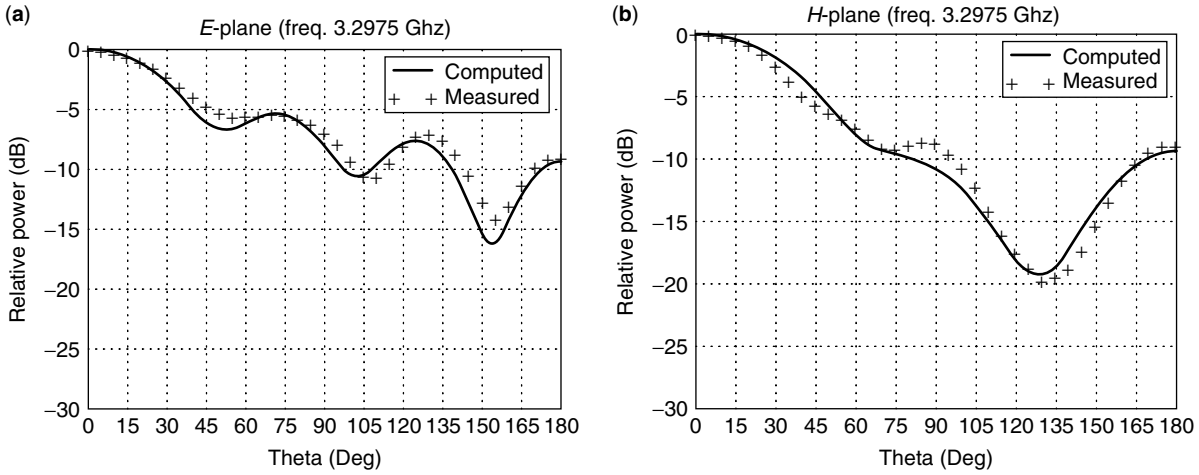


Figure 62. Predicted and measured *E*- and *H*-plane patterns of DFW.

BIOGRAPHIES

M. Hajian was born in Iran on April 21, 1957. He received his B.S. in Physics from the University of Oklahoma (USA) in 1981, and the M.S. degree in Electrical Engineering from Delft University of Technology (in the Netherlands) in 1990. Since 1990 he has been with the Microwave and Radar Laboratory of the Delft University of Technology. In 1995 he became a Senior Lecturer teaching a course on antennas. He is the Netherlands representative of EC/COST 260 on adaptive antennas. His major interests are antennas and propagation, smart antennnas, antenna near-field measurement techniques, and mobile communications systems.

L. P. Ligthart was born in Rotterdam, on September 15, 1946. He graduated with distinction in 1969 and received the M.S. degree in Electrical Engineering from Delft University of Technology. Since 1969 he has been with the Microwave Laboratory of the Delft University of Technology. In 1974 he became a Senior Lecturer teaching an undergraduate course on transmission line theory, antennnas, and propagation. From 1976 to 1977 he spent one year as a senior scientist at Chalmers University, Gothenburg in Sweden. In 1985 he received the Ph.D. degree in Technical Sciences based on his contributions in the design of miniaturized waveguide radiating elements.

Prof. Dr. Ligthart is Director of IRCTR, covering activities on antennas and propagation; radar, mobile, and satellite communication; remote sensing; and electromagnetic compatibility. His present interests include antennas and propagation, radar, and remote sensing.

He received the Vederprijs award in 1981, the IEE-Blumlein-Brown-Willams Premium Award in 1982, and the Doctor Honoris Cause from Moscow State Technical University of Civil Aviation in 1999. He is a fellow of IEE and IEEE, and the Netherlands representative of EC/COST 260 on adaptive antennas and EC/COST on advanced weather radar. He has published over 152 scientific papers.

BIBLIOGRAPHY

1. H.-C. Song et al., Four-branch single-mode waveguide power divider, *IEEE Photon. Technol. Lett.* **10**(12): 1760–1762 (Dec. 1998).
2. L. F. Libelo and C. M. Knop, A corrugated waveguide phase shifter and its use in HPM dual-reflector antenna arrays, *IEEE Trans. Microwave Theory Tech.* **43**(1): 31–35 (Jan. 1995).
3. T. Yoneyama, Millimeter-wave transmitter and receiver using the nonradiative dielectric waveguide, *IEEE MTT-S Digest* 1083–1086 (1989).
4. Y. T. Lo and S. W. Lee, *Antenna Handbook, Theory, Applications, and Design*, Van Nostrand Reinhold, New York, 1988, Chap. 24.
5. G. H. C. van Werkhoeven and A. K. Golshayan, Calibration aspects of the APAR antenna unit, *IEEE Trans. AP* **46**(6): 776–781 (June 1998).
6. A. B. Smolders, Design and construction of a broadband wide-scan angle phased array antenna with 4096 radiating elements, *IEEE-APS on Phased Array Systems and Technology*, Boston, 1996, pp. 87–92.
7. J. Bennett et al., Quadpack X-band T/R module for active phased array radar, *GAAS 98 Conf. Proc.*, Oct. 1998, pp. 63–67.
8. V. K. Lakshmeesha et al., A compact high-power S-band dual frequency, dual polarized feed, *Antennas and Propagation Society International Symposium*, Vol. 3 AP-S. Digest, 1991, pp. 1607–1610.
9. T. S. Bird, M. A. Sprey, K. J. Greene, and G. L. James, A circularly polarized X-band feed system with high transmit/receive port isolation, *Antennas and Propagation*, Vol. 1 Ninth International Conference on (Cof. Publ. No. 407), 1995 pp. 322–326.
10. P. Savi, D. Trinchero, R. Tascone, and R. Orta, A new approach to the design of dual-mode rectangular waveguide filters with distributed coupling, *IEEE Trans. Microwave Theory Tech.* **45**(2): 221–228 (Feb. 1997).
11. D. Crawford and M. Davidovitz, A 2-step waveguide E-plane filter design method using the semi-discrete finite element method, *IEEE Trans. Microwave Theory Tech.* **42**(7): 1407–1411 (July 1994).

12. C. A. Balanis, *Advanced Engineering Electromagnetic*, Wiley, 1989.

13. R. E. Collin, *Field Theory of Guided Waves*, IEEE Press, 1990.

14. M. Tian, P. D. Tran, M. Hajian, and L. P. Ligthart, Air-gap technique for matching the aperture of miniature waveguide antennas, *IEEE Instrumentation and Measurement Technology Conf.*, May 18–20, 1993, pp. 197–201.

15. M. Hajian, T. S. Lam, and L. P. Ligthart, Microstrip-to-waveguide transition for miniature dielectric-filled waveguide antenna, *Microwave Opt. Technol. Lett.* **12**(5): (Aug. 1996).

16. T. Q. Ho and Y.-C. Shih, Spectral-domain analysis of E-plane waveguide to microstrip transition, *IEEE-MTT* **37**(2): 388–392 (Feb. 1989).

17. Transition links waveguide and microstrip lines, *Microwaves RF* 119–120 (May 1994).

18. D. Li and R. Wang, Analysis of waveguide-to-microstrip transition, *Microwave Opt. Technol. Lett.* **5**(3): 128–130 (March 1992).

19. G. E. Ponchak and A. N. Downey, A new model for broadband waveguide-to-microstrip transition design, *Microwave J.* 333–343 (May 1988).

20. T. Q. Ho and Y.-C. Shih, Analysis of microstrip line to waveguide end launchers, *IEEE-MTT* **36**(3): 561–567 (March 1988).

21. P. M. Meaney, A novel transition from waveguide to microstrip, *Microwave J.* **33**(11): 145–148 (Nov. 1990).

22. B. N. Das and K. V. S. V. R. Prasad, Excitation of waveguide by stripline- and microstrip-line-fed slots, *IEEE-MTT* **34**(3): 321–327 (March 1986).

23. P. A. Rizzi, *Microwave Engineering*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

24. C. T. Tai, *Dyadic Green's Functions in Electromagnetic Theory*, Scranton Intext, Scranton, PA, 1971, Chap. 5, pp. 76–80.

25. T. Lam, M. Hajian, and L. Ligthart, *Excitation of MLA by Microstrip*, IRCTR internal thesis report, Aug. 1995.

26. M. Hajian, *Analysis and Design of Dielectric Filled Waveguide Antennas for Collision Avoidance Radar*, IRCTR internal report, Feb. 1995.

27. C. N. Capsalis, A rigorous analysis of a coaxial to shielded microstrip line transition, *IEEE-MTT* **37**: 1091–1098 (July 1989).

28. M. Tian, *Characterization of Miniature Dielectric Filled Open Ended Waveguide Antennas*, Ph.D. thesis, Delft Univ., Oct. 1995.

29. R. F. Harrington, *Time-Harmonic Electromagnetic Field*, McGraw-Hill, New York, 1961.

30. L. P. Ligthart, *Antenna Design and Characterization Based on the Elementary Antenna Concept*, Ph.D. thesis, Dutch Efficiency Bureau, 1985.

31. N. Marcuvitz, *Waveguide Handbook*, Dover, New York, 1965.

32. System Manual, HP 8510B, *HP Network Analyzer User's Guide* 1986.

33. J. C. Liberti, Jr. and T. S. Rappaport, *Smart Antennas for Wireless Communications: IS-95 and Third Generation CDMA Applications*, Prentice-Hall, 1999.

34. C. B. Dietrich, Jr., K. Dietze, J. R. Nealy, and W. L. Stutzman, Spatial, polarization, and pattern diversity for wireless handheld terminals, *IEEE Trans. on AP* **49**(9): 1271–1281 (Sept. 2001).

35. C. Passmann, G. Villino, and T. Wixforth, A polarization flexible phased array antenna for a mobile communication SDMA field trial, *Int. Microwave Symp. Digest*, Denver, June 8–13, 1997, Vol. 2, pp. 595–598.

36. S. R. Cloude and K. P. Papathanassiou, Polarimetric SAR interferometry, *IEEE Trans. Geosci. Remote Sens.* **36**(5): (Part 1) 1551–1565 (Sept. 1998).

37. Y. Liu, S. J. Frasier, and R. F. McIntosh, Measurement and classification of low-grazing-angle radar spikes, *IEEE Trans. Antennas Propag.* **46**(1): 27–40 (Jan. 1998).

38. R. F. M. Van den Brink, *Design a Miniaturized Feed at 4 GHz*, IRCTR internal thesis report, Aug. 1984.

39. C. A. Balanis, *Antenna Theory, Analysis and Design*, 2nd ed., Wiley, 1997.

40. A. R. Moumen, *Analysis and Synthesis of Compact Feeds for Large Multiple Beam Reflector Antennas*, Ph.D. thesis, Delft Univ., March 2001.

MILLIMETER-WAVE ANTENNAS

ZHIZHANG (DAVID) CHEN
Dalhousie University
Halifax, Nova Scotia, Canada

1. INTRODUCTION

1.1. Definition of Millimeter Waves — A Part of Electromagnetic Spectrum

As one of the key areas in information technology, telecommunications involve the transmission of electric signals that contain messages from one location to another location as well as processing of these signals. The carriers of the electrical signals are time-varying electromagnetic waves in the forms of electric current flow or radiowave propagation. Two parameters characterize a time-varying electromagnetic wave or signal. The first is called *frequency*, which is defined as the number of cycles of variations per second in unit of *hertz* (Hz). The second is called *wavelength* with units in meters. The wavelength describes the spatial period of repetition in space when a signal of a certain frequency travels in a medium. The relationship between a frequency f and a wavelength λ in free space is $f = c/\lambda$, where c is the speed of light. Theoretically, the frequency of an electromagnetic wave (or the electrical signal) can be from zero to infinity and the corresponding wavelength then goes from infinity to zero, leading to an electromagnetic spectrum of infinite extent.

Electromagnetic waves or electric signals of different frequencies have found many different applications due to their different characteristics. A zero-frequency signal is what we normally call DC (direct current). The most commonly used batteries produce DC energy. It is used in any battery-powered equipment such as CD players and flashlights. A 60-Hz signal is what is used in our electric power grid to distribute electric energy from power stations to our homes. Figure 1 illustrates different applications with

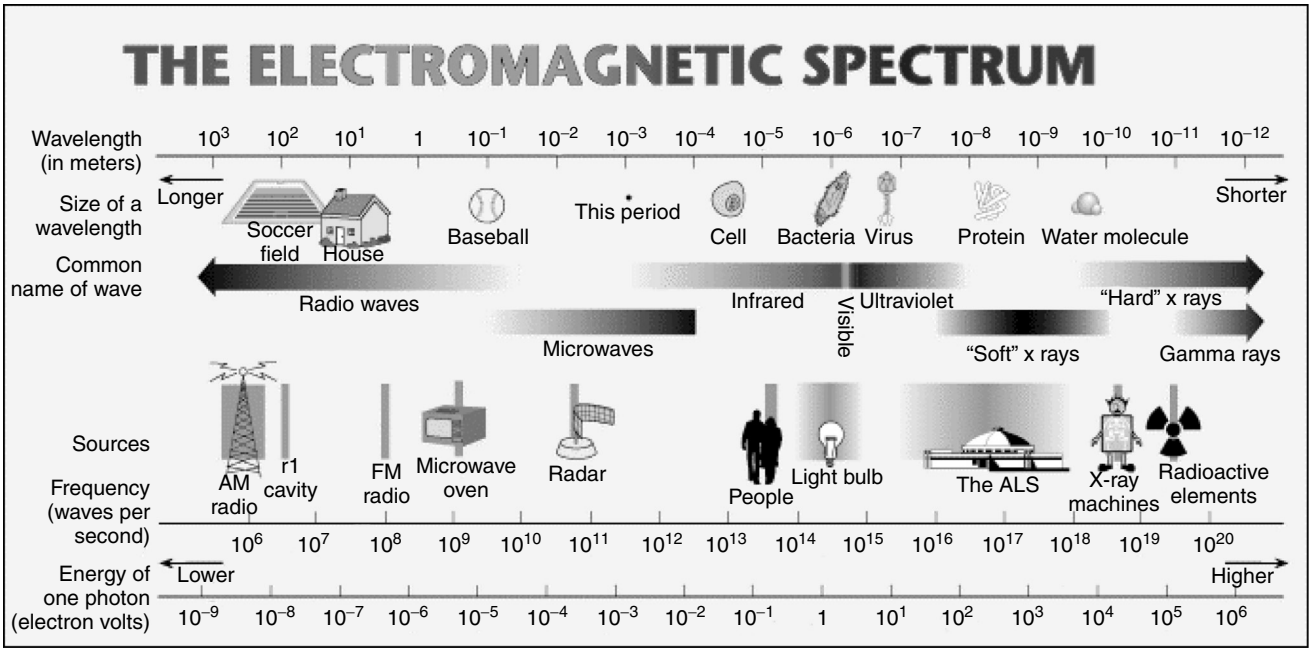


Figure 1. The electromagnetic spectrum (courtesy of the Advanced Light Source, Lawrence Berkeley National Laboratory).

Table 1. Band Classifications of the Radiofrequency Spectrum

Band Classification	Frequency f	Free-Space Wavelength λ	Applications
	<3 Hz	$>10^5$ (km)	Magnetotelluric sensing
Extremely low frequency (ELF)	3–30 Hz	10^5–10^4 km	Detection of buried metal objects
Superlow frequency (SLF)	30–300 Hz	10^4–10^3 km	Ionospheric sensing, electric power distribution, submarine communications
Ultralow frequency (ULF)	300 Hz–3 kHz	10^3–10^2 km	Audio signals on telephone
Very low frequency (VLF)	3–30 kHz	10–1 km	Navigation and position
Low frequency (LF)	30–300 kHz	1 km–100 m	Radio beacon, weather broadcast stations for air navigations
Medium frequency (MF)	300 kHz–3 MHz	100–10 m	AM broadcasting
High frequency (HF)	3–30 MHz	10–1 m	Shortwave broadcasting
Very high frequency (VHF)	30–300 MHz	1 m–10 cm	TV and FM broadcasting, mobile radio communication, air traffic control
Ultrahigh frequency (UHF)	300 MHz–3 GHz	10–1 cm	TV broadcasting, radar, radioastronomy, microwave ovens, cellular phones
Superhigh frequency (SHF)	3–30 GHz	1 cm–1 mm	Radar, satellite communication systems, aircraft navigation, radioastronomy, remote sensing
Extremely high frequency (EHF)	30–300 GHz	1–0.1 mm	Radar, advanced communication systems, remote sensing, radioastronomy

different frequency ranges of the electromagnetic spectrum. Note that contrary to what most people think, X-ray, infrared, or visible light are all parts of the electromagnetic wave spectrum with difference frequencies.

In theory, any part of the electromagnetic spectrum can be used for telecommunications. However, because of specific requirements for communications, most communication systems use the so-called radiofrequency spectrum where frequencies range from 3 Hz to 300 GHz (1 G = 10^9). For clarity, it is artificially divided into various bands, each being named in terms of its frequency and wavelength. Table 1 shows the classifications of the radio spectrum and their respective applications.

The millimeter-wave frequency band falls into the EHF band with frequencies ranging from 30 to 300 GHz (see Table 1). The term *millimeter waves* comes from the fact that the corresponding wavelength is in the range of millimeters in free space.

1.2. Applications of Millimeter Waves

As can be seen in Fig. 1, a millimeter wave is very close to the light in its spectrum position and possess very short wavelengths. Therefore, it has the properties similar to those of light, such as high resolution and large bandwidth. However, better than light, a millimeter wave experiences fewer environmental effects such as atmospheric absorption as it travels through the atmosphere. For this reason, millimeter waves, while possessing higher resolution and larger bandwidth than normal radio and microwave systems, can propagate through various transmission media. As the consequence of these properties, applications have been developed in areas of remote sensing/imaging, radioastronomy, plasma diagnosis, radar, and high-speed or broadband wireless and satellite communications. In particular, in telecommunications, in light of increasingly congested lower frequency bands and growing demands for high-speed data communications such as video transmissions, millimeter waves pose a very promising band of the radio spectrum to be utilized and have attracted growing attentions and research and development efforts.

1.3. Millimeter-Wave Antennas

One of the key components in any wireless telecommunication system is the antenna. It serves as the "eyes" and "ears" of a communication system, or technically, the interfacing system between the air and electronics. It radiates the radio signal it obtains from an electronic transmitter into space and sends the radio signals it detects in space to an electronic receiver. In general, to effect good radiation and detection, the size of an antenna must be proportional to the operating wavelength. The higher the operating frequency (or the shorter the wavelength), the smaller the antenna size. Roughly, the dimensions of an antenna are at least one-quarter to one-half of the wavelengths. The example is the antenna tower at an AM radiobroadcasting station. Because of the operating wavelengths of the AM signals are normally in the range of hundreds of meters, the antenna heights have to be tens or even hundreds of meters. In contrast, the millimeter-wave antennas tend to be very small and in the range of centimeters and millimeters, because millimeter waves have very short wavelengths. Such a feature, as well as the potential broad bandwidths with millimeter waves, has made millimeter-wave antennas very attractive for future short-range and high-speed communication systems. Figure 2 shows the millimeter-wave antennas made of traditional conical and pyramidal horn antenna structures. When designed to operate at 90–140 GHz, the dimensions for the conical horn are about 13 mm (in diameter) × 28 mm (in length) and for the pyramidal horn 20 mm (in width) × 15 mm (in height) × 41 mm (in length).

Like any radio antennas, the important specifications for the millimeter-wave antennas are gain, radiation pattern, return loss, bandwidth, and efficiency [1]. The *gain* of an antenna describes the degree of radio energy concentration by an antenna in a direction, in reference to the isotropic antenna that radiates the energy uniformly in all the directions. The *radiation pattern* of an antenna describes the radiation power intensity distribution along all the directions. The *return loss* of an antenna indicates the degree of the energy reflected by the antenna due to the impedance mismatch between the antenna and a transmitter or a receiver. The *bandwidth* is defined as the frequency range within which the performance of the antenna, with respect to some characteristics (e.g., return loss or gain), conforms to a specified standard. *Efficiency* takes into account the energy loss due to imperfections of conductors and substrates used.

Figure 2. Millimeter-wave horn antennas (courtesy of QuinStar Technology Inc).

In general, millimeter-wave (mm) antennas can be categorized into five groups in terms of their structures and configurations: (1) the traditional reflector and lens antennas, (2) waveguide-based antennas, (3) printed-circuit antennas, (4) active integrated antennas, and (5) optically controlled and integrated antennas. Figure 3 lists the categories and the various antenna types under each category.

It should be noted that Fig. 3 is simply a general presentation of the techniques used so far in developing millimeter-wave antennas. The groupings presented are not absolutely clearcut among various millimeter-wave antennas reported so far. An antenna may belong to two or three groups simultaneously. For instance, an active integrated antenna may also belong to the group of printed-circuit antennas as it may be fabricated on a planar structure.

2. REFLECTOR AND LENS ANTENNAS

The traditional reflector and lens antennas are still being used for millimeter-wave applications because of their simplicity in operational principles and constructions. A typical reflector, illustrated in Fig. 4, consists of a feed that illuminates the conducting reflector with radio millimeter waves. The reflector surface is shaped in such a way that the radio fields scattered by the reflector will illuminate the area in a desired pattern. For instance, a parabolic surface will focus the radio energy in one direction, while the others will focus the radio energy in a certain coverage area. A shaped reflector antenna for 60-GHz indoor wireless networks has been reported [2]. It achieved a very good circular coverage with edge illumination less

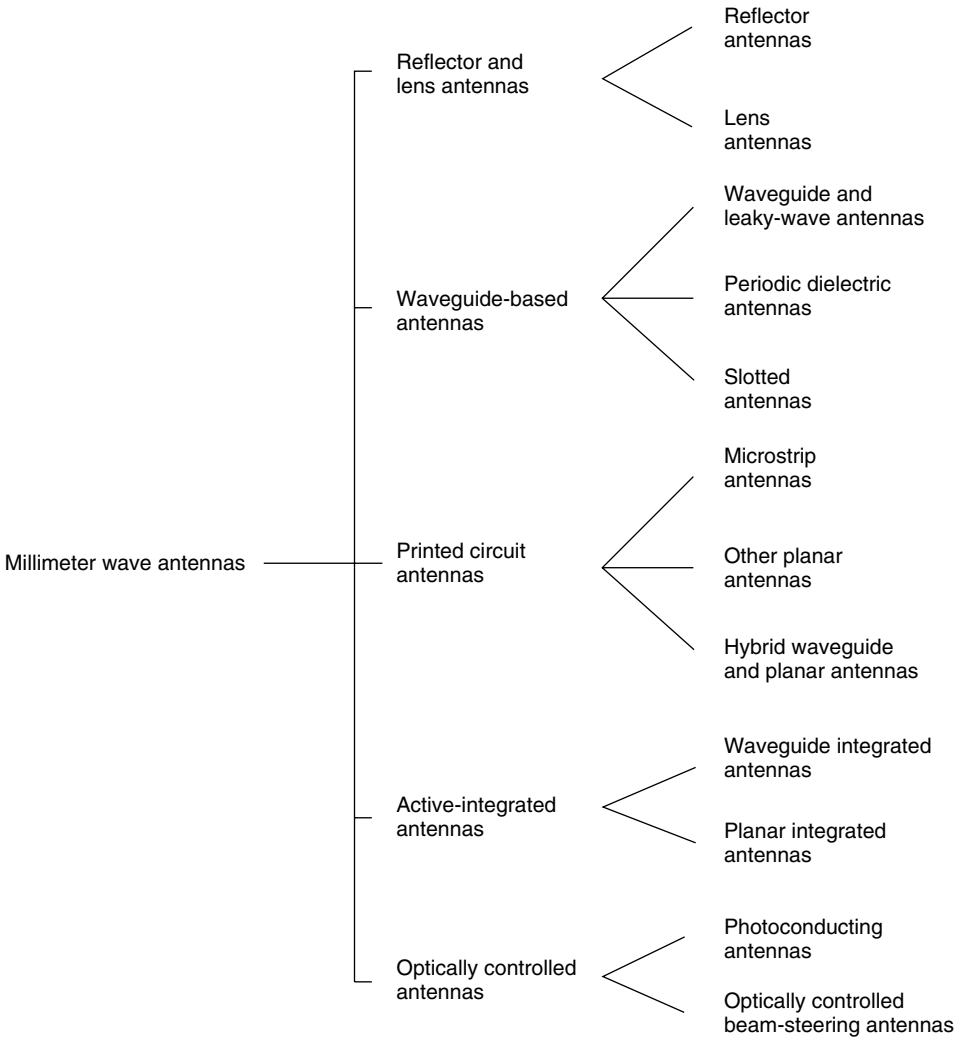


Figure 3. Division of millimeter-wave antennas in terms of structure.

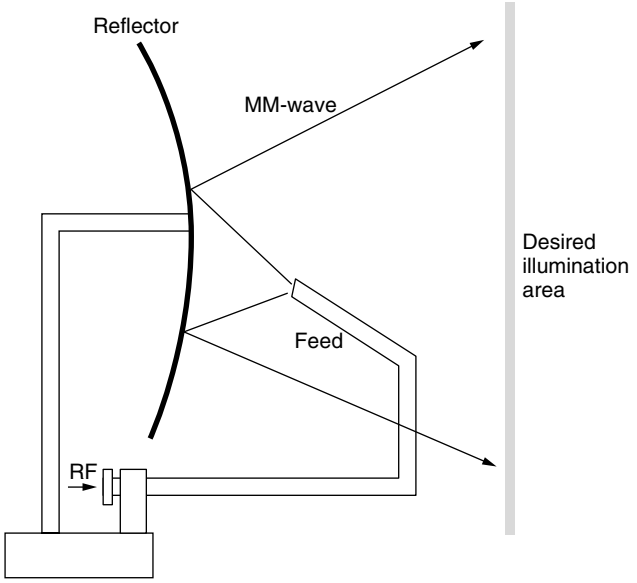


Figure 4. A shaped reflector antenna (the shape of the reflector is designed in such a way that the waves reflected by the reflector only illuminate the area desired).

than −10 dB of the desired boresight illumination. The size of the antenna is, however, rather bulky. The reflector has a diameter of 30 cm.

A millimeter-wave lens antenna is illustrated in Fig. 5. The lens is formed by a low-loss dielectric, and its surface is designed in such a way that the waves coming from the radiating patch will be diffracted at the lens–air interface into the air in the desired angle. When multiple radiating elements are placed at the different locations, a multibeam antenna can be achieved. It was reported that such an antenna achieved a directivity of 25.9 dB at 30 GHz with a beamwidth of 6.6° [3]. The diameter of the lens is ∼50 mm (∼5 cm), and the height of the whole structure is ∼80 mm (∼8 cm).

3. WAVEGUIDE-BASED ANTENNAS

As their name implies, waveguides are devices that guide and transmit microwave and millimeter (mm)-wave energy from one point to another. Traditional waveguides include coaxial, rectangular and circular waveguides, which are simply hollow metal tubes with cross-sections of two concentric circles, a rectangle and a circle (see

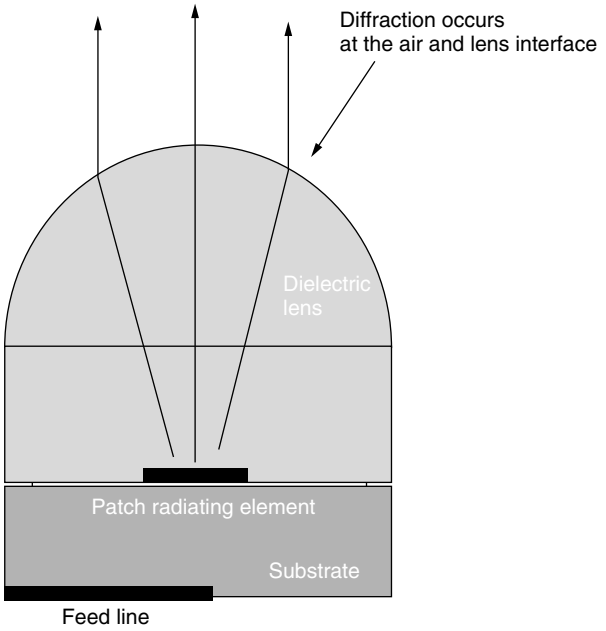


Figure 5. A lens millimeter-wave antenna (the shape of the lens is designed in such a way that the fields diffracted at the lens-air interface will radiate in the desired angles) (redrawn from Fig. 1 of Ref. 3, © 2001 IEEE).

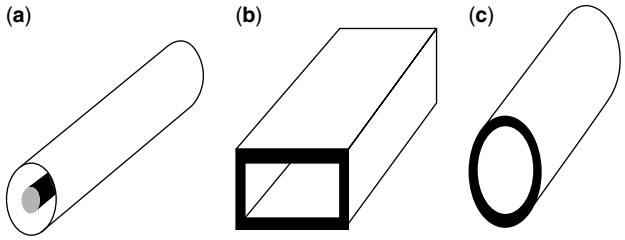


Figure 6. Traditional waveguides that are made of hollow metal tubes with different cross-sectional shapes sections: (**a**) coaxial line; (**b**) rectangular waveguide; (**c**) circular waveguide.

Fig. 6). The more modern waveguides, or guided-wave structures, are mostly planar structures that facilitate the integration with integrated circuits. They include striplines, microstrip lines, and coplanar lines as shown in Fig. 7. It has been proved theoretically and experimentally that when the shapes of the cross sections are uniform along the guides (i.e., do not vary along the longitudinal direction), microwave and mm-wave energy will propagate without incurring much energy leakage and radiation. However, when the uniformity of the cross sections is perturbed, energy leakage or radiation will occur. A simple example is to leave a waveguide abruptly cut open to air or free space. The microwave and mm-wave energy will then radiate into space. Consequently, by intentionally introducing perturbations in the wavequides along their longitudinal directions, radiation into space can be achieved in a controlled and desired manner. A specific advantage of these waveguide-based antennas is their compatibility with the waveguides from which they derived the energy, thus facilitating integrated designs with the waveguide structures. A few of these types of antennas are introduced below.

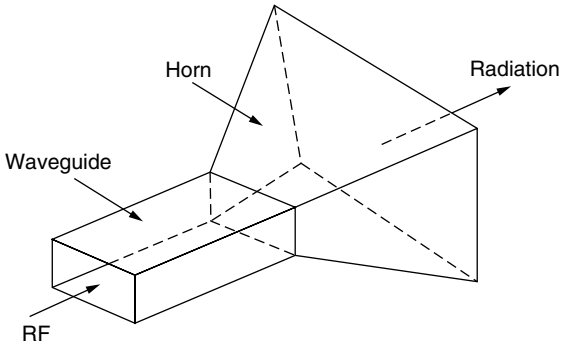


Figure 8. A pyramidal horn antenna.

3.1. Waveguide-Derived Antenna

The first type of antenna is the *waveguide-derived antennas*, where the waveguide structures are perturbed at their ends. They include horn antennas and leaky-wave antennas.

In a *horn antenna*, a longitudinally uniform waveguide is made with an open end. To ensure that most of energy in a waveguide radiates efficiently out into space, a transition from the waveguide to the open end is required. A typical transition is a horn-shape extension that transforms a small aperture of the waveguide to a large aperture (see Fig. 8). The radiation pattern of such an antenna is normally end-fire type. The gain of a horn antenna runs from 20 to 40 dB. The efficiency in light of the conducting loss and spillover energy can be as high as 85%. The length of the horn is less than 10 cm, and the width and height are less than 8 cm.

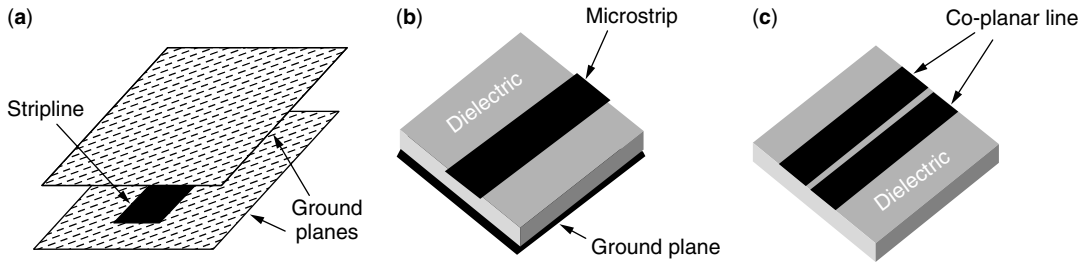


Figure 7. The planar guided wave structures with all the metal strips are planar: (**a**) striplines; (**b**) microstrip line; (**c**) coplanar lines.

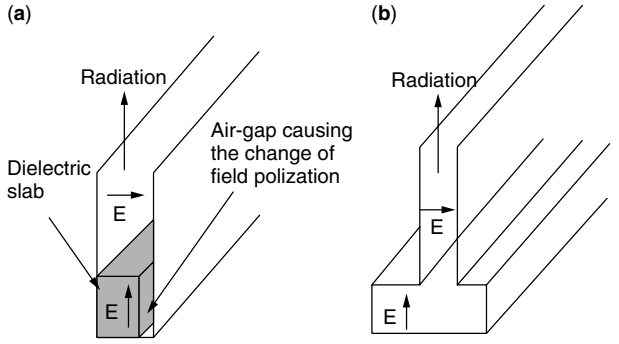


Figure 9. Two leaky-wave antennas (fields in the main guides are perturbed so that the fields in the upper arms propagate without cutoff and then radiate): (**a**) NRD guide antenna; (**b**) special groove guide antenna (redrawn from Fig. 2 of Ref. 18, © 1992 IEEE).

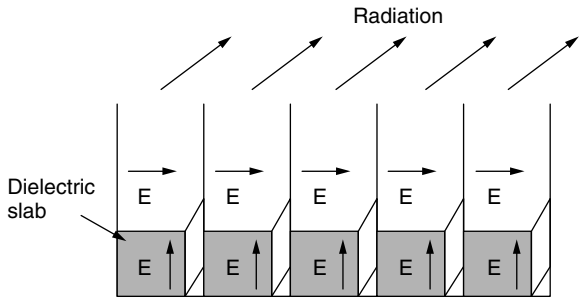


Figure 10. Array of leaky-wave NRD guide antennas (redrawn from Fig. 4 of Ref. 18, © 1992 IEEE).

Figure 9 shows two typical leaky-wave waveguide antennas, one with the use of a nonradiative dielectric (NRD) waveguide and the other with a special groove waveguide [4]. Both waveguides are bisected horizontally relative to their unperturbed waveguides with extended upper arms and closed at the bottom. Horizontally polarized fields will be produced in the foreshortening upper arms as a result of diffraction of the fundamental modes around the asymmetric air gaps or open aperture. The fields will then propagate without cutoff frequencies in a transverse electromagnetic parallel-plate mode to the end of the arms and radiate outward. Figure 10 shows an array of NRD waveguide antennas. It is typically 10–50 wavelengths long. By varying the phases to the feed systems for the antenna elements, the radiation beam can be steered in the longitudinal plane.

3.2. Periodic Dielectric Antenna

The second type of waveguide-based antenna is the *periodic dielectric antenna* [5–7], which normally consists of a uniform dielectric waveguide with a periodic surface perturbation that may take the form of a dielectric grating or a metal grating (see Fig. 11). The structure is designed in such a way that the fundamental mode is excited in the nonperturbed section of the waveguide and then transformed, as a result of the grating in the perturbed area, into a leaky wave that radiates into space. It has been shown theoretically and experimentally that as the frequency is increased, the main-beam direction scans from backfire, through broadside and into the forward quadrant. There exists, however, a possible stopband where an internal resonance occurs as a result of the periodic structures. Such a resonance will inhibit radiation. Fortunately, in most cases, such a stopband is narrow.

3.3. Slotted Waveguide Antenna

The third type of a waveguide-based antenna is a *slotted waveguide antenna*, where slots are opened on the sidewalls of the waveguide (see Fig. 12). As the waves propagate down the waveguide, fields radiate through the slots out into space. Since the number of slots can be large, the gain can be made as high as 35 dB with efficiency of 75% at 60 GHz [8].

4. PRINTED-CIRCUIT ANTENNAS

Microstrip and other printed-circuit antennas are planar types of antennas that are fabricated on multiplayer dielectric substrate structures. They are simple in structure, easy to fabricate by lithography, and convenient for circuit integration. Most of them are low-profile, lightweight, and low-cost devices potentially conformal to any planar surfaces. The most commonly seen are microstrip patch/dipole antennas, planar slotted antennas, and hybrid waveguide–planar antennas.

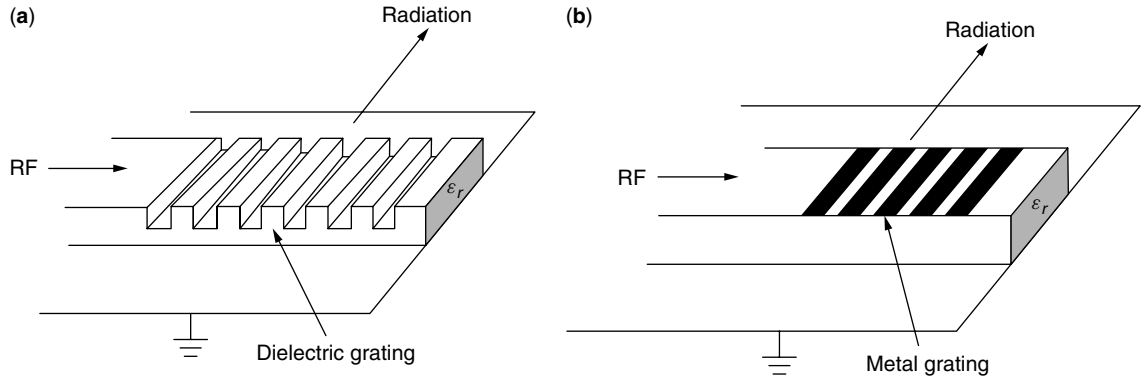


Figure 11. The periodic dielectric antennas (redrawn from Fig. 1 of Ref. 18, © 1992 IEEE).

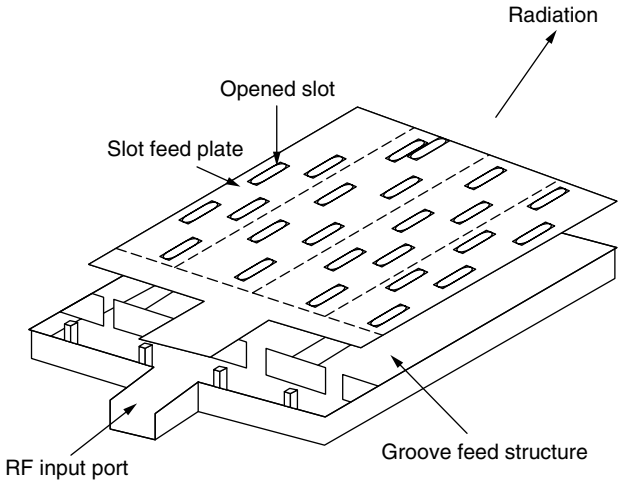


Figure 12. A single-layer waveguide slot antenna array (redrawn from Fig. 3 of Ref. 8, © 1997 IEEE).

The *microstrip patch antennas* are perhaps the simplest antennas, formed either by a patch or a planar dipole (see Fig. 13). The patch or the planar dipole serves as the radiating element. In general, the bandwidth of a microstrip antenna is quite narrow (e.g., <5%) because most of the fields are trapped between the patches and the ground planes. There are two other problems peculiar to mm-wave applications: fabrication tolerance and losses primarily associated with the feed systems.

Because of the small wavelengths, absolute fabrication tolerance for fabricating mm-wave lines is very small. For instance, for the width of a feedline, the order of a few tenths of a millimeter is required for operating frequencies of 30–100 GHz. In addition, conducting losses are relatively high and efficiency is potentially low because of the high frequencies of millimeter waves and the small cross section of a feedline. All these factors have limited the application of microstrip antennas to narrowband applications up to frequencies of 140 GHz [9,10].

To resolve the difficulties encountered in the microstrip antennas, other planar antennas have been proposed. One of them is the top-loaded coplanar waveguide fed aperture stacked patch antenna (Fig. 14), in which the co-planar transmission line is used as the feedline but is coupled to the radiation elements by means of aperture coupling. Two radiating patches resonating at two neighboring frequencies are stacked on top of each other with dielectric substrates in between. As a result, the bandwidth is increased from <1% to >15% [11].

A combination of waveguide structures with printed-circuit antennas was also been proposed. One of them is a two-dimensional printed dipoles with dual polarizations suspended in pyramidal horns (see Fig. 15). The horn is etched into a silicon wafer structure and the dipoles are printed by photolithographic techniques. The radiation characteristics of the antennas are determined by both the horn structures and the dipoles. Such a structure facilitates the integration with planar circuits while maintaining the efficiency and high gain of the horn antennas [12].

5. ACTIVE INTEGRATED ANTENNAS

The term *active integrated antenna* (AIA) refers to a class of radiating structures where the radiating elements are not only used for radiation but also serve as integral parts of active components such as resonators and filters. As a result, conventional 50-Ω feedline connections between radiating elements and active components and subsequent impedance matching elements are no longer necessary. In other words, both radiating elements and active components can be integrated and fabricated on the same substrate or multiplayer substrates. This leads to the obvious advantages of compactness, reliability, reproducibility, and potential low-cost.

A variety of AIA structures have been reported. It is difficult to comprehensively review all these antennas. Nevertheless, an attempt is made here to divide AIA structures into two groups in terms of the structures and configurations: waveguide-derived structures and planar structures.

Figure 16 shows a monolithic integrated waveguide single-slot mixer on a GaAs substrate mounted in a TE_{10} waveguide and a horn. RF signals received by the horn will be coupled to the GaAs active diode chip for signal

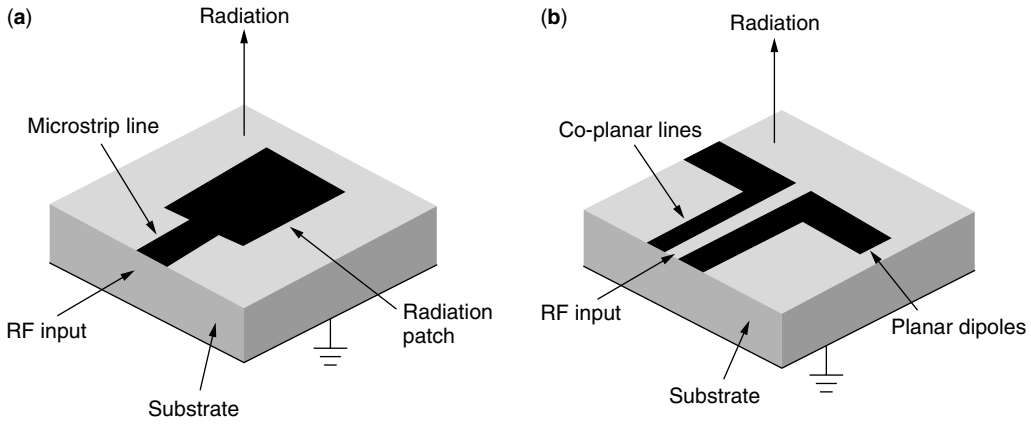


Figure 13. Microstrip patch (**a**) and planar dipole (**b**) antennas.

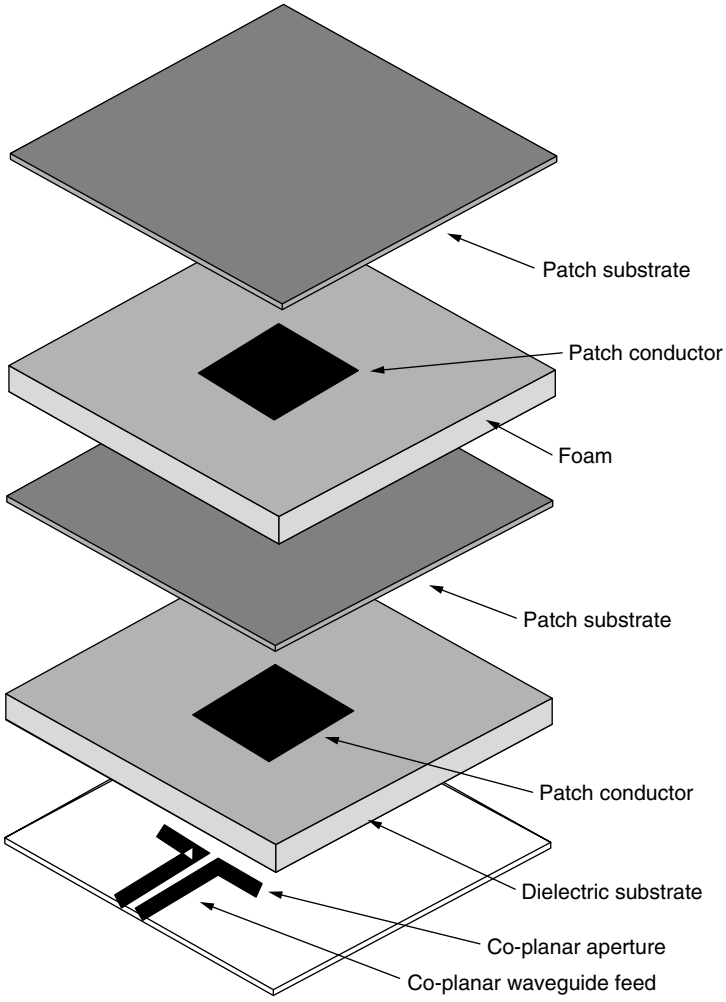


Figure 14. The stacked microstrip patch antenna (the radiating patch resonates at neighboring frequencies to increase the bandwidth) (redrawn from Fig. 2 of Ref. 11, © 2000 IEEE).

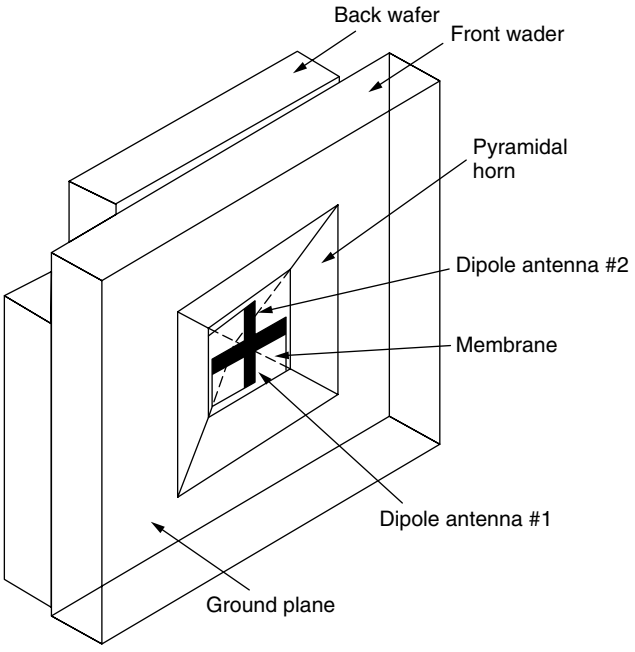


Figure 15. An integrated horn with dual polarizations for integrated balanced mixers (redrawn from Fig. 30 of Ref. 19, © 1992 IEEE).

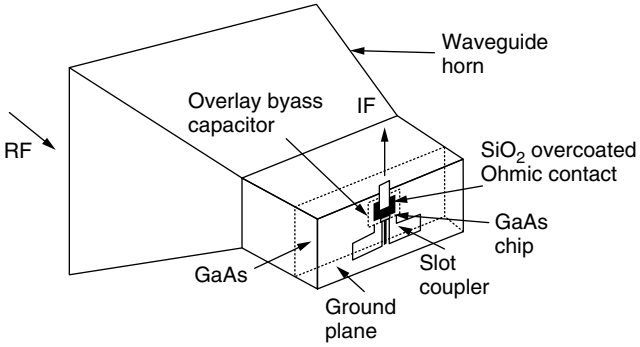


Figure 16. Monolithic integrated circuit single-slot mixer on a GaAs substrate in a TE_{10} waveguide (redrawn from Fig. 6 of Ref. 19, © 1992 IEEE).

mixings. The IF signal is directly output from the back of the waveguide horn [13].

Figure 17 illustrates a conceptually typical planar active integrated antenna. The radiating elements are a periodic, linear, series-fed microstrip patch array that also serves as the resonator to the oscillator. The field-effect transistor (FET) is used to improve the DC-to-RF conversion efficiency. The power output of the FET is

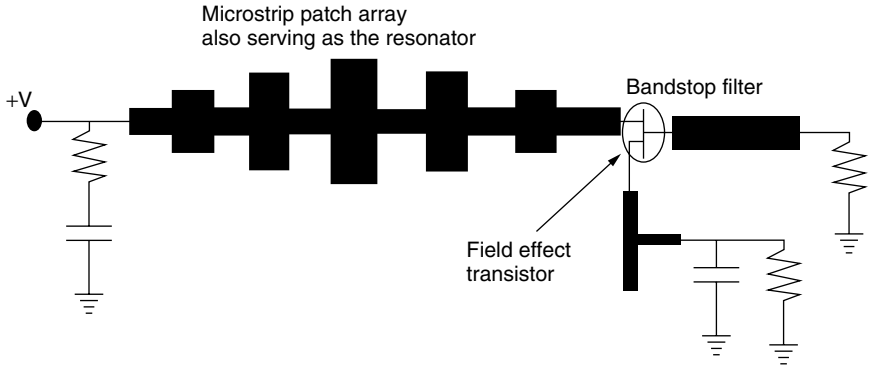


Figure 17. Periodic microstrip patch array with integrated FET (redrawn from Fig. 9 of Ref. 18, © 1992 IEEE).

delivered to the antenna that radiates in the broadside direction. The gate of the integrated FET is terminated in a bandstop filter that provides the correct reactance at the oscillating frequency [14]. A slight modification with additions of RF input circuitry could lead to a transceiver circuit, where the FET performs the dual functions, as the source for the transmitted signal and as self-oscillating mixer for downconversion of the received signal.

6. OPTICALLY CONTROLLED ANTENNAS

By integrating optical circuits and components with conventional millimeter-wave structures, optically controlled millimeter-wave antennas have been developed since the early 1990s. In these antenna structures, the properties of mm-wave radiations are controlled by lasers or optical signals. There are two groups of such antennas: (1) photoconducting antennas that result in generating millimeter waves and (2) optically controlled beam-steering antennas.

In the first case (see Fig. 18) the antenna consists of a planar dipole, a photoconductor deposited in between the feed gap of the dipole, coplanar strip transmission line, and contact pads for photoconductor biasing. The antenna is excited by illuminating the photoconductor with optical pulses, and the millimeter wave is generated as the result of the illumination and dipole resonance. By modulating the bias applied to the photoconductor, modulated millimeter waves can also be obtained [15].

In the second case, the antenna array is developed for beam scanning. The phases of RF signals fed to each array element are controlled with an optical means, either with photoconductors or with optical wavelength-dependent time-delay dispersive structures. Two examples are shown in Figs. 19 and 20. In Fig. 19, the semiconductor slab is illuminated with a special pattern formed by photomasks, creating a photoinduced plasma grating on the slab (that behaves similarly to a metal grating). The millimeter waves that couple to the slab through the dielectric waveguide will then interact with the plasma grating and radiate out of the slab in a specific direction. The direction is dependent on the grating pattern that is controlled by the photomasks [16]. In Fig. 20, millimeter-wave signals are first modulated onto an optical signal and split into four modulated optical signals. These four signals propagate through an optically controlled dispersive prism and thus have different time delays. They are demodulated with photodiodes and fed to antenna arrays, leading to beam steering that is dependent on the phase delays of the four signals. It was reported that such an antenna achieved squint-free steering across $\pm 60^{\circ}$ azimuthal span and over the entire Ka band (26.5–40 GHz) [17].

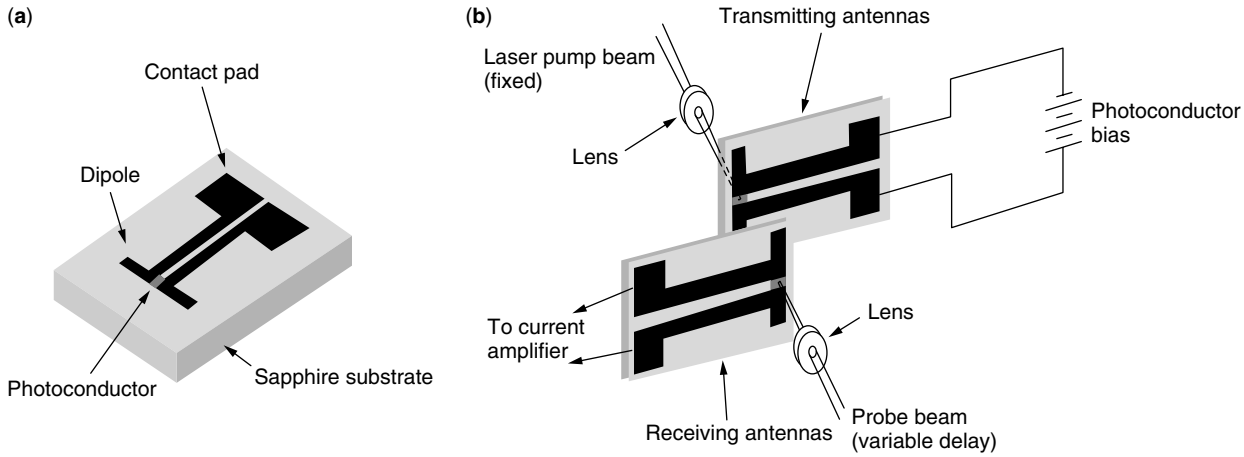


Figure 18. Photoconducting antenna structure (**a**) for generating electric short pulses that contain (**b**) millimeter-wave and submillimeter-wave components (redrawn from Figs. 1 and 2 of Ref. 15, © 1988 IEEE).

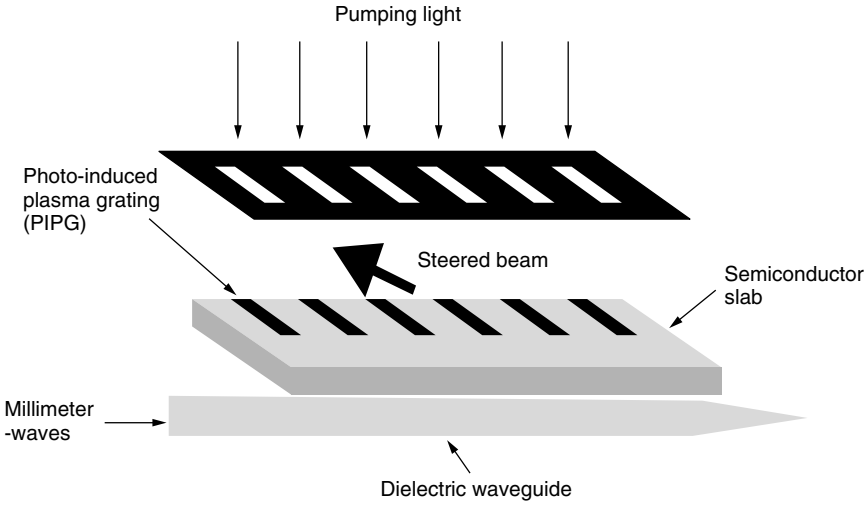


Figure 19. The beam-steering antenna with photoinducing plasma grating (redrawn from Fig. 1 of Ref. 16, © 1997 IEEE).

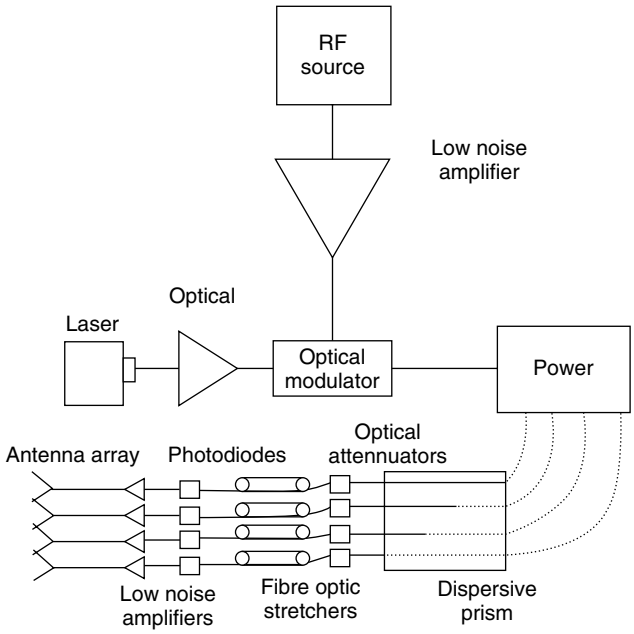


Figure 20. The fiberoptic beam-steering antenna (redrawn from Fig. 1 of Ref. 17, © 2001 IEEE).

7. FUTURE TRENDS OF MILLIMETER-WAVE ANTENNAS

Because of their high frequencies, millimeter waves are a promising medium for broadband and high-speed wireless communications. Their short wavelength permits circuits and systems to be small and compact. Their near-light properties allow for precise imaging and remote sensing in different environments such as clouds and damp air. As a result, millimeter-wave circuits and systems have been investigated and studied extensively since the early 1980s, including millimeter-wave antennas.

As described above, tremendous progress in the understanding and fabricating millimeter-wave antennas have been achieved [18,19]. Numerous antenna structures have been developed to meet different application requirements and to achieve better performance. In particular, active integrated antennas and optically controlled antennas have been attracting much attention.

Nevertheless, like the design of other millimeter-wave circuits and systems, the main challenges associated with the design of the millimeter-wave antennas are (1) low gains and low powers offered by active components, (2) relatively high conductor losses and substrate losses, and (3) the cost of components and systems. All these factors have the limiting effects in applications of millimeter-wave circuits and systems. It is expected that these three challenges will also continue to be the topics of interest of future research and development of millimeter-wave antennas.

BIOGRAPHY

Zhizhang (David) Chen received his B. Eng. degree in radio engineering in 1982 from Fuzhou University, P. R. China, his M. A. Sc. degree in radio engineering in 1986 from Southeast University, P. R. China, and his Ph.D. degree in electrical engineering in 1992 from the University of Ottawa, Ottawa, Ontario, Canada. He was a lecturer with the Department of Radio Engineering, Fuzhou University from 1985 to 1988, and has held a Natural Science and Engineering Research Council postdoctoral fellowship with the Department of Electrical and Computer Engineering, McGill University, Montreal, Québec, Canada, from January of 1993 to August of 1993. Since September of 1993, he has been with the Department of Engineering, Dalhousie University, (formerly Technical University of Nova Scotia), Halifax, Nova Scotia, Canada, where he is presently an associate professor. Dr. Chen has published over 100 refereed journal/conference papers and industrial reports in the areas of RF/microwave circuit and system design and computational electromagnetics. His general research areas are in RF/microwave engineering and applied electromagnetics for communications and microelectronics. His current teaching and research interests include RF/microwave CAD, RF interconnection & packaging, antenna design, numerical modeling and simulation, and wireless circuit and system design.

BIBLIOGRAPHY

1. C. A. Ballanis, *Antenna Theory*, 2nd ed., Wiley, 1997.
2. P. F. M. Smulder, S. Khusial, and M. H. A. J. Herben, A shaped reflector antenna for 60 GHz indoor wireless LAN access points, *IEEE Trans. Vehic. Technol.* **50**(2): 584–592 (March 2001).
3. X. Wu, G. V. Eleftheriades, and T. E. V. Deventer-Perkins, Design and characterization of single- and multiple-beam mm-wave circularly polarized substrate lens antennas for wireless communications, *IEEE Trans. Microwave Theory Tech.* **49**(3): 431–441 (March 2001).
4. F. K. Schwering and A. A. Oliner, Millimeter-wave antennas, in K. Chang, ed., *Handbook of Microwave and Optical Components*, Wiley, New York, 1988, Chap. 12.2.
5. T. Itoh and B. Adelseck, Trapped image guide leaky-wave antennas for millimeter-wave applications, *IEEE Trans. Antennas Propag.* **AP-30**(5): 505–509 (May 1982).
6. F. Schwering and S. T. Peng, Design of dielectric grating antennas for millimeter-wave applications, *IEEE Trans. Microwave Theory Tech.* **MTT-31**: 199–209 (Feb. 1983).
7. M. Guglielmi and A. A. Oliner, A practical theory for image guide leaky-wave antennas loaded by periodic metal strips, *Proc. 17th Eur. Microwave Conf.*, Rome, Italy, Sept. 11–17, 1987, pp. 549–554.
8. M. Ando and J. Hirokawa, Novel single high-gain and high-efficiency single-layer slotted waveguide arrays in 60 GHz band, *Proc. 10th Int. Conf. Antennas and Propagation*, Edinburg, UK, April 14–17, 1997, pp. 464–668.
9. F. K. Schwering and A. A. Oliner, Millimeter-wave antennas, in Y. T. Lo and S. W. Lee, eds., *Antenna Handbook*, Van Nostrand Reinhold, New York, 1988, Chap. 17.
10. M. A. Weiss, Microwave antennas for millimeter waves, *IEEE Trans. Antennas Propag.* **AP-29**: 171–174 (Jan. 1981).
11. W. S. T. Rowe and R. B. Waterhouse, Comparison of broad-band millimeter-wave antenna structures for MMIC and optical device integration, *Digest of 2000 IEEE Int. Antennas and Propagation Symp.*, Salt Lake City, UT, 2000, pp. 1390–1393.
12. W. Y. Ali-Ahmad and G. M. Rebeiz, 94 GHz integrated horn monopulse antennas, *IEEE Trans. Antennas Propag.* **39**(7): 820–825 (July 1991).
13. B. J. Clifton, G. D. Alley, R. A. Murphy, and I. H. Mroczkowski, High-performance quasioptical GaAs monolithic mixers at 110 GHz, *IEEE Trans. Electron Devices* **28**: 135–157 (Feb. 1981).
14. J. Birkeland and T. Itoh, FET-based planar circuits for quasi-optical sources and transceivers, *IEEE Trans. Microwave Theory Tech.* **37**(9): 1452–1459 (Sept. 1989).
15. P. R. Smith, D. H. Auston, and M. C. Nuss, Subpicosecond photoconducting dipole antennas, *IEEE J. Quant. Electron.* **24**(2): 255–260 (Feb. 1988).
16. V. A. Manasson, L. S. Sadovnik, V. A. Yepishin, and D. Marker, An optically controlled MMW beam-steering antenna based on a novel architecture, *IEEE Trans. Microwave Theory Tech.* **45**(8): 1497–1500 (Aug. 1997).
17. D. A. Tulchinsky and P. J. Matthews, Ultrawide-band fiber-optic control of a millimeter-wave transmit beamformer, *IEEE Trans. Microwave Theory Tech.* **49**(7): 1248–1253 (July 2001).
18. F. K. Schwering, Millimeter-wave antennas, *Proc. IEEE* **80**(1): 92–102 (Jan. 1992).
19. G. M. Rebeiz, Millimeter-wave and terahertz integrated circuit antennas, *Proc. IEEE* **80**(11): 1748–1770 (Nov. 1992).

MILLIMETER WAVE PROPAGATION

EDWARD E. ALTSHULER
Electromagnetics Technology Division
Hanscom AFB, Massachusetts

1. INTRODUCTION

The millimeter wave region of the electromagnetic spectrum generally covers wavelengths in the range from about 2 cm down to 1 mm (15–300 GHz). These limits are based on wavelength and on the nature and magnitude of the interaction between the wave and the atmosphere. The propagation characteristics of electromagnetic waves in this region are of particular interest because the waves have a strong interaction with lower atmospheric gases and particulates. Although the interaction with the atmosphere does not change abruptly at these limits, it does become weaker at longer wavelengths and stronger at shorter wavelengths. Thus the concepts presented here can generally be extended to either slightly longer or slightly shorter wavelengths. We shall often refer to the "window regions" of the millimeter wave spectrum. These are considered the low-attenuation regions between the gaseous absorption resonances. In particular, wavelengths between the 1.35-cm water vapor resonance and the 5-mm oxygen resonance, the 5- and 2.5-mm oxygen resonances, and the 2.5-mm oxygen resonance and 1.6-mm water vapor resonance compose the window regions. Low-attenuation regions also exist at wavelengths longer than 1.35 cm and shorter than 1.6 mm.

The physics of the interaction between millimeter waves and the atmosphere is extremely complex, so many facets are considered beyond the scope of this article. However, an effort is made to provide the reader with a general understanding of the mechanisms of this interaction; if in-depth details are required, they can be obtained from the references. Likewise, complicated mathematical expressions are used only to illustrate concepts that are considered important and cannot be satisfactorily explained otherwise. Many of the figures contain information that is "typical" or "average", that is, it is intended to provide the reader with an estimate of the magnitude of an interaction. More quantitative results are available from the cited references. In this article we first review the physics of the interaction of millimeter waves with the atmosphere and then describe the effects of the atmosphere on both terrestrial and earth–space communications.

1.1. Propagation Effects

Atmospheric gases and particulates often have a profound effect on millimeter waves and thus limit the performance

of many millimeter wave systems. Because the densities of these gases and particulates generally decrease with altitude, the effects of the atmosphere on the propagated wave are strongest very close to the earth and tend to diminish at higher altitudes. For this reason millimeter waves propagating above the tropopause — the altitude at which the temperature remains essentially constant with increasing height (~10–12 km) — are assumed to be unaffected. For the clear atmosphere there are essentially two types of interaction. The stronger interaction is the absorption–emission produced by oxygen and water vapor. The other interaction takes place with the refractivity structure of the atmosphere, which is often divided into two categories: gross structure and fine structure. For the gross structure it is assumed that the atmosphere is a horizontally stratified continuum characterized by a refractivity that normally decreases slowly with increasing altitude and is wavelength-independent; that is, it affects all wavelengths from microwaves through millimeter waves in the same way. For the refractivity fine structure, the atmosphere is viewed as an inhomogeneous medium consisting of small pockets of refractive index that vary both temporally and spatially. Because these pockets have different sizes, this interaction is wavelength-dependent. Although the terrain is not actually part of the atmosphere, it does interface with the atmosphere and can affect millimeter wave propagation. Thus, multipath propagation produced by the terrain and diffraction by prominent obstacles on the earth's surface are also reviewed. Atmospheric particulates range in size from micrometers to close to 1 cm in diameter. For particles very small compared to wavelength, the only significant propagation effects are those of absorption and emission. As the particles become larger with respect to wavelength, scattering effects become pronounced. In Section 2 we shall review the effects of the clear atmosphere and then of atmospheric particulates on millimeter wave propagation. We show that although the effects of the clear atmosphere are generally not as severe as those due to atmospheric particulates, they cannot be disregarded, even in the window regions and especially at short millimeter wavelengths.

1.2. Applications

Potential applications of millimeter waves have been considered for many years [1]. However, for most applications, atmospheric effects have always imposed limitations on system performance. The principal applications of millimeter waves have been in the areas of communications and radar. Probably the first application for which millimeter waves were considered was communications. The discovery of the circular electric waveguide mode TE_{01}, in the late 1930s prompted the use of these wavelengths for a waveguide communication system, because waveguide attenuation for that mode decreases with decreasing wavelength. It is believed that the most significant contribution resulting from this effort was not so much the development of the system itself but rather the research that was directed toward a whole new line of millimeter wave equipment and techniques required for this application namely, sources, amplifiers, detectors, and waveguide components. Through the years consideration was often given to utilizing millimeter waves for point-to-point communications. However, proposed systems never materialized, principally because they were not economically competitive with those at longer wavelengths; they also lacked reliability because of atmospheric effects and inferior components.

The need for new types of communication systems and the need to alleviate increasing spectrum congestion finally led to a reappraisal of millimeter waves. The availability of large bandwidths makes this region of the spectrum particularly attractive for high-data-rate earth–space communication channels. Furthermore, high-gain, high-resolution antennas of moderate size and lightweight compact system components are indeed applicable for space vehicle instrumentation. Millimeter waves provide an excellent means for obtaining secure communication channels. For satellite–satellite links, where all propagation is above the absorptive constituents of the lower atmosphere, narrow-beamwidth antennas may be operated at a wavelength where atmospheric attenuation is very high (i.e., $\lambda \sim 5$ mm); thus the signal is confined by the antenna to a narrow cone and then absorbed by the lower atmosphere before it reaches the earth. Another application for secure communications is ship-to-ship and short terrestrial links; in these cases attenuation is sufficiently high at millimeter waves to allow a detectable signal only over short distances. Finally, point-to-point radio-relay systems that were previously not considered feasible are now in operation. More recent studies have shown that attenuation in the lower atmosphere can be combatted using very short hops and diversity techniques; also, satisfactory system performance can now be obtained with solid-state components quite economically.

Because the beamwidth of an aperture is inversely proportional to wavelength, antennas having the same size aperture, have better resolution at millimeter wavelengths than at longer wavelengths. Furthermore, because range resolution is a function of bandwidth, improved range resolution is also possible at the shorter wavelengths.

2. ATMOSPHERIC EFFECTS ON PROPAGATED WAVES

Atmospheric gases and particulates may severely alter the properties of millimeter waves. For the clear atmosphere the most pronounced effect is absorption due to the gases, oxygen, and water vapor. However, refraction, scattering, diffraction, and depolarization effects are also reviewed, and it is shown that under special conditions their impact on the propagated wave can be significant. For an atmosphere containing particulates, the absorption and scattering by rain seriously limit propagation at millimeter wavelengths. However, the effects of smaller, water-based particulates are also significant, particularly at shorter millimeter wavelengths.

2.1. Clear Atmosphere

We shall consider an atmosphere "clear" if it is free of atmospheric particulates. Thus the only interaction that takes place is that between the propagated wave and

the atmospheric gases (and terrain). We shall see that for very low elevation angles the gross structure of the refractive index causes the propagated wave to be bent and delayed, whereas the fine structure of the refractive index scatters the propagated wave and may produce scintillations. The gases, oxygen, and water vapor absorb energy from the wave and reradiate this energy in the form of noise. Finally, if the wave is propagating close to the earth's surface, part of the energy may be reflected from the surface, or it may be diffracted by prominent obstacles on the surface and then interfere with the direct signal. If some form of reflection or scattering takes place, the propagated wave may also become depolarized. All of these effects are discussed in detail in Sections 2.1.1–2.1.5.

2.1.1. Refraction. The lower atmosphere is composed of about 78% nitrogen, 21% oxygen, and 1% water vapor, argon, carbon dioxide, and other rare gases. The densities of all these gases, except for water vapor, decrease gradually with height; the density of water vapor, on the other hand, is highly variable. The index of refraction of the atmosphere is a function of the temperature, pressure, and partial pressure of water vapor. Because the refractive index, n, is only about 1.0003 at the earth's surface, it is often expressed in terms of a refractivity N, where

$$N = (n - 1) \times 10^6 \tag{1}$$

The radio refractivity can be approximated by the following theoretical expression, which has empirically derived coefficients [2]

$$N = \frac{77.6P}{T} + \frac{3.73 \times 10^5 e}{T^2} \tag{2}$$

where P is the atmospheric pressure (in millibars), T the absolute temperature (in degrees kelvin), and e is the water vapor pressure (in millibars). This expression consists of two terms; the first is often referred to as the "dry" term, and the second as the "wet" term, because it is a function of water vapor. Equation (2) neglects dispersion effects and in principle does not hold near absorption lines; however, in practice it is assumed valid down to a wavelength of ~3 mm and is often considered acceptable down to wavelengths of even 1 mm, particularly in the window regions. The refractivity decreases approximately exponentially with height and is often expressed as

$$N(h) = N_s e^{-bh} \tag{3}$$

where h is the height above sea level (in kilometers); N_s, the surface refractivity; $b = 0.136\ \text{km}^{-1}$; $N(h)$ is the refractivity at height h (in kilometers).

As mentioned earlier, the gross behavior of atmospheric refractivity affects millimeter waves in much the same way as microwaves; because effects such as bending, time delay, and ducting are amply treated elsewhere [3], they will only be summarized here. The refractive index fine-scale structure is too complex to be treated by simple refraction theory. Because of the variability in the scale sizes of the refractive inhomogeneities, the effects are wavelength-dependent, and thus the interaction is different for millimeter waves than for microwaves.

2.1.1.1. Bending. The angular bending is due primarily to the change of the index of refraction of the atmosphere with height. Because the refractivity generally decreases with height, the wave passing obliquely through the atmosphere is bent downward. Thus the apparent elevation angle of a target tends to appear slightly higher than its true elevation angle, and this difference is called the *angle error*. For a horizontally stratified atmosphere, the angle error is zero at zenith and increases very slowly with decreasing elevation angle. This error becomes appreciable at low elevation angles, and for a standard atmosphere it approaches a value of about 0.7° near the horizon. For illustration we plot in Fig. 1 the angle errors of targets at altitudes of 90 km and infinity (radio source) for an atmosphere with a surface refractivity of $313N$ units that decreases exponentially with height. For a typical atmosphere the refractivity decreases at a rate of about $40N$ units/km near the surface and then more slowly at higher altitudes; this produces superrefraction. As the gradient of this decrease in refractivity increases, the wave is bent more toward the earth's surface, and when the gradient decreases at a rate of $157N$ units/km, the wave travels parallel to the surface; this condition is called *ducting*. For still steeper gradients the wave will actually be bent into the earth.

There are times when the decrease in refractivity is less than $40N$ units/km and the wave is bent less than normal toward the earth; this is called *subrefraction*. In actuality we have a curved ray passing over a curved earth. It is sometimes easier to visualize a straight ray over a curved earth (or a curved ray over a flat earth).

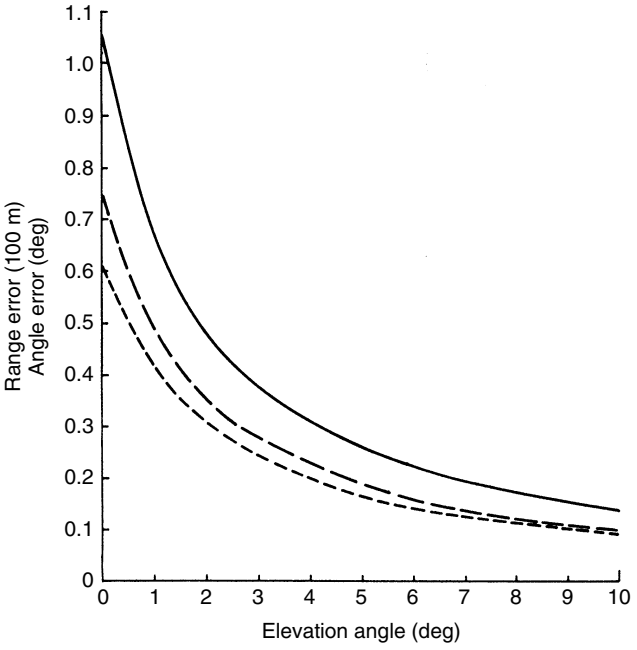


Figure 1. Typical tropospheric refraction angle and range errors:———, range error; — —, angle error ($h = \infty$);- - - - , angle error ($h = 90$ km).

For a vertical negative gradient of $40N$ units/km, it can be shown that an earth with an effective radius about four-thirds that of the true earth with the wave propagating in a straight line is equivalent to the true case — the curved ray path over the curved earth. The ratio of the effective earth radius to the true earth radius is often designated k. If the refractivity gradient is less than $40N$ units/km, then k decreases and reaches unity for the case of no gradient. As the gradient becomes more positive, k approaches zero. When the negative gradient is more negative than $40N$ units/km, k is larger than $\frac{4}{3}$ and approaches infinity for the special case of ducting, for which the gradient reaches $157N$ units/km. For still larger negative gradients, k becomes negative, because the flat earth for $k = \infty$ becomes curved in the opposite sense.

2.1.1.2. Time Delay. Time delay occurs primarily because the index of refraction of the atmosphere is greater than unity, thus slowing down the wave, and to a lesser extent because of the lengthening of the path by angular bending. For navigation systems, the range is determined from time-delay measurements; thus the additional time delay produced by the troposphere results in a corresponding range error. This error causes the target to appear farther away than its true distance. Navigation systems such as the Global Positioning System (GPS) must correct for this range error [4]. For a typical atmosphere the range error is slightly larger than 2 m in the zenith direction and increases very slowly with decreasing elevation angle. The range error becomes much larger at lower elevation angles, and for a standard atmosphere it approaches a value of about 100 m near the horizon. For illustration, we plot in Fig. 1 the range error for an atmosphere with a surface refractivity of $313N$ units that decreases exponentially with height.

2.1.1.3. Refraction Corrections. As seen in Fig. 1, the angle and range errors become appreciable for very low elevation angles. It has been shown that both errors are strongly correlated with the surface refractivity, and for many applications adequate corrections based on a linear regression on N are possible [3,4]. More accurate corrections can be obtained by actually measuring the vertical refractivity profile and then calculating the corrections. The principal limitation of this approach is that a horizontally stratified atmosphere is usually assumed, and this is not always valid for the long distances traversed at low elevation angles. It has been shown that the range error is correlated with the brightness temperature of the atmosphere, and techniques to take advantage of this dependence have been proposed [5–8]. One of the most effective methods for obtaining angle error corrections involves the use of "targets of opportunity." These may be either calibration satellites or radio sources, the angular positions of which are normally known to an accuracy of the order of microradians. In principle, the angular error of the calibration source is measured before the target is tracked. If the target is in the same general direction as the calibration source and the atmospheric refractivity does not change appreciably with time, then the correction can be determined directly. For range error corrections, differential GPS uses a known location on the surface as a reference.

2.1.1.4. Scintillations. So far we have discussed the effects of the gross refractivity structure of the atmosphere on the propagated wave. The atmosphere also has a fine-scale refractive index structure, which varies both temporally and spatially and thus causes the amplitude and phase of a wave to fluctuate; these fluctuations are often referred to as *scintillations* [9–11]. The refractive index structure is envisioned to consist of pockets of refractive inhomogeneities that are sometimes referred to as "turbulent eddies" and may be classified by size into three regions: the input range, the inertial subrange, and the dissipation range. The two boundaries that separate these regions are the outer and inner scales of turbulence, L_0 and l_0, respectively. These are the largest and smallest distances for which the fluctuations in the index of refraction are correlated. A meteorologic explanation of how these pockets are generated is beyond the scope of this article; however, in simple terms, large parcels of refractivity, possibly of the order of hundreds of meters in extent, continually break down into smaller-scale pockets. These pockets become smaller and smaller until they finally disappear. The very large pockets in the input range have a complex structure, and at the present time there is no acceptable formulation of the turbulence properties of this region. Pockets having scale sizes of less than $\sim$1 mm have essentially no turbulent activity, and for all practical purposes the spectrum of the covariance function of the refractive index fluctuations, $\phi_n(k)$, equals zero. The inertial subrange bounded by L_0 and l_0 has a spectrum

$$\varphi_n(k) = 0.033\, C_n^2\, k^{-11/3} \tag{4}$$

for $2\pi/L_0 < k < 2\pi/l_0$, where C_n is the structure constant and k the wavenumber (not to be confused with the ratio k of the effective earth radius to the true earth radius used earlier). The phase fluctuations arise from changes in the velocity of the wave as it passes through pockets of different refractive indices. As the wavelength becomes shorter, the changes in phase increase proportionally. The amplitude fluctuations arise from defocusing and focusing by the curvature of the pockets.

2.1.2. Absorption and Emission. Atmospheric gases can absorb energy from millimeter waves if the molecular structure of the gas is such that the individual molecules possess electric or magnetic dipole moments. It is known from quantum theory that, at specific wavelengths, energy from the wave is transferred to the molecule, causing it to rise to a higher energy level; if the gas is in thermal equilibrium, it will then reradiate this energy isotropically as a random process, thus falling back to its prior energy state. Because the incident wave has a preferred direction and the emitted energy is isotropic, the net result is a loss of energy from the beam. The emission characteristics of the atmosphere may be represented by those of a blackbody at a temperature that produces the same emission; therefore the atmospheric emission is often expressed as an apparent sky temperature.

Because absorption and emission are dependent on the same general laws of thermodynamics, both are expressed in terms of the absorption coefficient. Using Kirchhoff's law and the principle of conservation of energy, one can derive the radiative transfer equation, which describes the radiation field in the atmosphere that absorbs and emits energy. This emission is expressed as

$$T_a = \int_0^\infty T(s)\gamma(s) \exp\left(-\int_0^\infty \gamma(s')\,ds'\right) ds \qquad (5)$$

where T_a is the effective antenna temperature, $T(s)$ is the atmospheric temperature, $\gamma(s)$ is the absorption coefficient, and s is the distance from the antenna (ray path). In simpler terms

$$T_a = T_m(1 - e^{-\gamma s}) \qquad (6)$$

where T_m is the atmospheric mean absorption temperature within the antenna beam. Solving for the attenuation, we obtain

$$A = \gamma s = 10 \log\left(\frac{T_m}{T_m - T_a}\right) \qquad (7)$$

where A is in decibels. The only atmospheric gases with strong absorption lines at millimeter wavelengths are water vapor and oxygen. The absorption lines O_3, CO, N_2O, NO_2, and CH_2O are much too weak to affect propagation in this region.

2.1.2.1. Water Vapor. The water vapor molecule has an electric dipole moment with resonances at wavelengths of 13.49, 1.64, and 0.92 mm (22.24, 183.31, and 325.5 GHz) in the millimeter wave region. In general, the positions, intensities, and linewidths of these resonances agree well with experimental data. There are, however, serious discrepancies between theoretical and experimental absorption coefficients in the window regions between these strong lines; experimental attenuations are often a factor of 2–3 times larger than theoretical values.

Although the cause of the discrepancy is not known, indications are that either the lineshapes do not predict enough absorption in the wings of the resonances or there is an additional source of absorption that has not yet been identified. It should be mentioned that there are over 1800 water vapor lines in the millimeter wave/infrared spectrum, 28 of which are at wavelengths above 0.3 mm. Because the wings of these lines contribute to the absorption in the window regions, very small errors in the lineshapes could significantly affect the overall absorption. In an effort to overcome this problem, several workers have introduced an empirical correction term to account for the excess attenuation [12]. In addition to the uncertainty of the absorption coefficient of water vapor, there is also the problem of water vapor concentration. The amount of water vapor in the lower atmosphere is highly variable in time and altitude and has densities ranging from a fraction of a gram per cubic meter for very arid climates to 30 g/m^3 for hot and humid regions; for this reason it is very difficult to model. A plot of the water vapor absorption as a function of frequency is shown in Fig. 2 for a density of 7.5 g/m^3. Because the attenuation, α, is linearly proportional to the water vapor density, except for very high concentrations, attenuations for other water vapor densities are easily obtained.

2.1.2.2. Oxygen. The oxygen molecule has a magnetic dipole moment with a cluster of resonances near a wavelength of 5 mm (60 GHz) and a single resonance at 2.53 mm (118.75 GHz). Although the >30 lines near a wavelength of 5 mm are resolvable at low pressures (high altitudes), they appear as a single pressure-broadened line near sea level owing to a large number of molecular collisions. Even though the magnetic dipole moment of oxygen is approximately two orders of magnitude weaker than the electric dipole moment of water vapor, the net absorption due to oxygen is still very high, simply because it is so abundant. The fact that the distribution of oxygen throughout the atmosphere is very stable makes it easy to model. A plot of oxygen attenuation as a function of frequency is shown in Fig. 2 along with that of water vapor. Note the importance of water vapor attenuation at very short wavelengths.

2.1.3. Scattering. The principal effect of the pockets of refractive index on the propagated wave is to produce scintillations, as described in Section 2.1.1. When the pockets are of the order of a wavelength in size, they can also scatter the signal. At millimeter wavelengths and for line-of-sight paths, this scattered field is generally very weak compared to the direct signal and is not considered

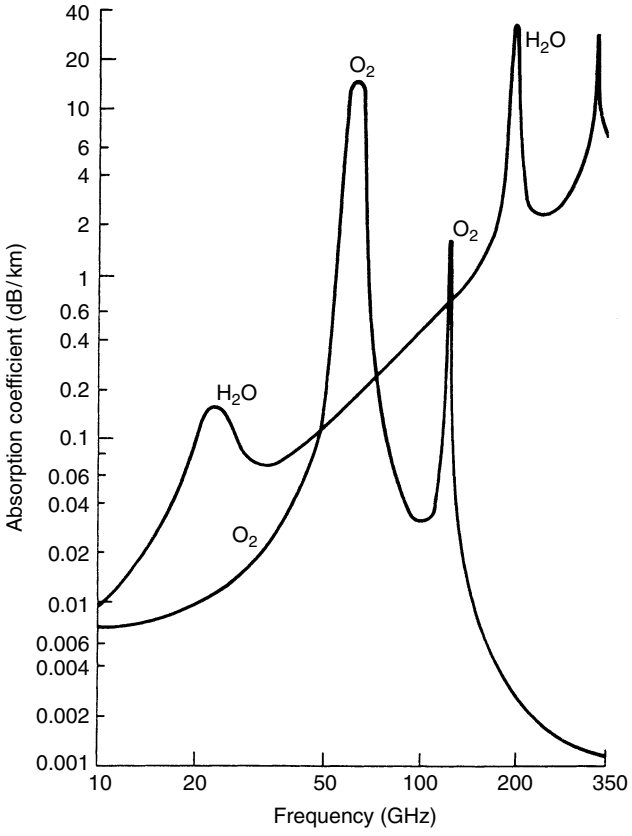


Figure 2. Absorption coefficients for water vapor and oxygen.

significant. Electromagnetic waves scattered from the earth's surface may interfere with the direct signal; this is called *multipath propagation*. The extent of multipath is dependent on the geometry of the transmitter and receiver with respect to the surface, their respective beamwidths and polarizations, and the dielectric constant and surface roughness of the terrain. Let us first consider a surface that is relatively smooth with respect to wavelength. The reflection coefficients of vertically and horizontally polarized waves for a nonmagnetic surface are

$$\Gamma_v = \frac{\varepsilon \sin\alpha - (\varepsilon - \cos^2\alpha)^{1/2}}{\varepsilon \sin\alpha + (\varepsilon - \cos^2\alpha)^{1/2}} \tag{8}$$

$$\Gamma_h = \frac{\sin\alpha - (\varepsilon - \cos^2\alpha)^{1/2}}{\sin\alpha + (\varepsilon - \cos^2\alpha)^{1/2}} \tag{9}$$

where α is the grazing angle and $\varepsilon = \varepsilon' - j\varepsilon''$ is the complex dielectric constant. For very small grazing angles the magnitudes of the reflection coefficients approach unity and the phases approach 180°. As the grazing angle increases, the magnitude and phase of the vertically polarized wave fall off sharply and those of the horizontally polarized wave decrease very slightly. The sharp falloff of the vertically polarized reflection coefficient can be explained as follows. For most surfaces, with the exception of very dry ground, $|\varepsilon| \gg 1$. With this approximation, Eq. (8) can be rewritten as

$$\Gamma_v = \frac{(\varepsilon)^{1/2}\sin\alpha - 1}{(\varepsilon)^{1/2}\sin\alpha + 1} \tag{10}$$

Note that the numerator is $((\varepsilon)^{1/2}\sin\alpha - 1)$. Thus at some angle the numerator approaches zero; this is the Brewster angle, and if the terrain were a perfect dielectric [$(\varepsilon)^{1/2}$ is real], then the reflection coefficient would actually go to zero. As the grazing angle approaches normal incidence the vertical reflection coefficient increases and finally equals the horizontally polarized reflection coefficient at normal incidence. When the surface is relatively smooth, the scattering is predominantly specular; that is, it can be considered coherent. As the surface becomes rougher, a diffuse, incoherent component appears, and for a very rough surface the scattered signal is predominantly diffuse. The criterion usually applied for characterizing surface roughness is that introduced by Rayleigh. It is based on the phase difference of adjacent rays reflected from a rough surface; when the path difference between these rays increases to about 90°, the surface is assumed to transform from smooth to rough. Obviously this transition is very gradual and should be interpreted as such.

Mathematically the surface can be considered smooth when $h\sin\alpha < \frac{1}{8}\lambda$, where h is the height of a surface irregularity. It must be emphasized that even at millimeter wavelengths, for which λ is very small, typical surfaces tend to look smooth at very low grazing angles. We have reviewed the general characteristics of multipath propagation. Now let us summarize multipath propagation in the context of millimeter waves. At longer wavelengths the reflection coefficient of a vertically polarized wave is significantly lower than that of a horizontally polarized wave, particularly in the vicinity of the Brewster angle, so microwave systems are often designed to operate with vertical polarization to minimize multipath interference. At millimeter wavelengths this polarization dependence is of less importance, because the reflection coefficients of vertically and horizontally polarized waves are comparable for most millimeter wave applications. First, multipath effects at these short wavelengths will generally occur only at very small grazing angles, because most surfaces based on the Rayleigh criterion appear rough for larger grazing angles. Furthermore, because the dielectric constants of most surfaces tend to remain constant or decrease with decreasing wavelength, the Brewster angle increases and the reflection coefficient of the vertically polarized wave does not drop off as rapidly with increasing grazing angle. Therefore at millimeter wavelengths multipath is confined to much lower grazing angles than at microwave wavelengths and is thus less sensitive to polarization. As the surface becomes rough with respect to wavelength, the grazing angle must become very small to have specular reflection.

2.1.4. Diffraction. Electromagnetic waves incident on an obstacle may be bent around that obstacle; this is known as *diffraction*. The extent of diffraction is dependent on the shape and composition of the obstacle, its position with respect to the direct path of the incident wave, and the wavelength. Tradition diffraction theory has been used to treat simple shapes such as individual knife edges, rounded edges, and in some instances sets of these edges. An underlying assumption is that the knife edge is very sharp or the rounded edge very smooth with respect to wavelength. It is also often assumed that the edge is a perfect conductor, although solutions have been obtained for edges having finite conductivity. Diffraction loss is often expressed as a function of the dimensionless Fresnel parameter v, which is, in turn, a function of the geometric parameters of the obstacle and path. For knife-edge diffraction the Fresnel parameter can be defined as

$$v = h\left[\frac{2}{\lambda}\left(\frac{1}{d_1} + \frac{1}{d_2}\right)\right]^{1/2} \tag{11}$$

where h is either the height of the obstacle above the direct path or the distance of the obstacle below the path and d_1 and d_2 are the respective distances of transmitter and receiver from the knife edge. For illustration, let us assume that the knife edge is midway between transmitter and receiver; then $d_1 = d_2 = \frac{1}{2}d$ and

$$|v|^2 = \frac{8h^2}{\lambda d} \tag{12}$$

where v is positive when the ray path is below the edge and negative when the ray path is above the edge. It is known that the diffraction loss is approximately zero for $v < -3$ and very high for $v > 3$, so $-3 < v < 3$ can be considered the region of interest.

Equation (12) can be expressed as

$$h = (\tfrac{1}{8}\lambda d v^2)^{1/2} = \tfrac{1}{2}v(\tfrac{1}{2}\lambda d)^{1/2} \tag{13}$$

Because d is generally on the order of kilometers and λ on the order of millimeters, h can be only on the order of meters. Thus, from a practical standpoint only isolated obstacles such as small hills or buildings would produce diffraction effects at millimeter wavelengths.

2.1.5. Depolarization. Depolarization of an electromagnetic wave can occur when the incident wave is scattered and a cross-polarized component is produced along with the copolarized component. It is defined as

$$|\text{ depolarization }| = 20 \log \left(\frac{|E_x|}{|E_y|} \right) \tag{14}$$

where E_x and E_y are the cross-polarized and copolarized components, respectively, and the depolarization is in decibels. Olsen [13] has summarized in detail both the mechanisms that can produce depolarization during clear air conditions and some experimental observations of this phenomenon. He divides these mechanisms into two groups: those that are independent of the cross-polarized pattern of the antenna (a perfect plane-polarized wave) and those that are dependent on the cross-polarized pattern. In principle, depolarization can arise from scattering by refractive inhomogeneities or from terrain. For the plane-polarized wave it appears that depolarization due to refractive multipath is insignificant but that depolarization due to terrain multipath can be much stronger [13]. For an antenna having a measurable cross-polarized pattern, it is believed that both atmospheric and terrain multipath mechanisms contribute to depolarization of the wave. Although most experimental results of depolarization by the clear atmosphere have been obtained at centimeter wavelengths, there is no reason to believe that the same effects will not occur at millimeter wavelengths. However, because both atmospheric and terrain multipath may normally be weaker at millimeter wavelengths than at microwave wavelengths, the depolarization may not be as severe.

2.2. Atmospheric Particulate Effects

In this section we discuss the degrees of absorption, scattering, and depolarization that may occur from atmospheric particulates. We shall see that rain is by far the most important of the particulates, for two reasons: (1) the interaction of rain with millimeter waves is very strong and (2) rain occurs more often than do other particulates. Thus the interaction between rain and millimeter waves is discussed in detail.

2.2.1. Absorption and Scattering. Millimeter waves incident on atmospheric particulates undergo absorption and scattering; the degree of each is dependent on the size, shape, and complex dielectric constant of the particle and the wavelength and polarization of the wave. The following Mie expression can be used for calculating the absorption and scattering from a dielectric sphere:

$$Q_t = \frac{\lambda^2}{2\pi} \text{ Re } \sum_{n=1}^{\infty} (2n+1)(a_n^s + b_n^s) \tag{15}$$

where Q_t represents losses due to both absorption and scattering, and a_n^s and b_n^s are complicated spherical Bessel functions that correspond to the magnetic and electric modes of the particle, respectively. Q_t has the dimension of area and is usually expressed in square centimeters. Physically, if a wave with a flux density of S (W/cm^2) is incident on the particle, then $S \times Q_t$ is the power absorbed or scattered. When the circumference of the particle is very small compared to wavelength (i.e., $\pi D \ll \lambda$), then the scattering and absorption losses can be represented by

$$Q_s = \left(\frac{\lambda^2}{2\pi} \right) \left(\frac{4}{3\rho^6} \right) \left| \frac{n^2 - 1}{n^2 + 2} \right|^2 \tag{16}$$

and

$$Q_a = \left(\frac{\lambda^2}{2\pi} \right) (2\rho^3) \text{ Im } \left| -\frac{(n^2 - 1)}{n^2 + 2} \right| \tag{17}$$

where $\rho = kD/2 = \pi D/\lambda \ll 1$ and n is the complex index of refraction. Because ρ is very small, the loss due to scattering, which is proportional to ρ^6, will be much smaller than that due to absorption, which is proportional to ρ^3. This condition is often referred to as the Rayleigh approximation, for which

$$Q_s \propto \frac{1}{\lambda^4} \quad \text{and} \quad Q_a \propto \frac{1}{\lambda}$$

Because the scattering loss is often assumed negligible, the total loss is proportional to the volume of the drop. Often the backscatter cross section (or radar cross section) is of interest:

$$\sigma = \left(\frac{\lambda^2 \rho^6}{\pi} \right) \left| \frac{n^2 - 1}{n^2 + 2} \right|^2 = \frac{3}{2} Q_s \tag{18}$$

The relationship with Q_s arises because Rayleigh scatterers are assumed to have the directional properties of a short dipole and the directivity of a dipole in the backscatter direction is 1.5 times that of an isotropic source. As the drop becomes large with respect to wavelength, the Rayleigh approximation becomes less valid and the Mie formulation in Eq. (16) must be used.

2.2.2. Depolarization. Atmospheric particulates having a nonspherical shape will depolarize a wave (produce a cross-polarized component) if the major or minor axis of the particulate is not aligned with the E field of the incident wave. The extent of the depolarization is a strong function of the size, shape, orientation, and dielectric constant of the scatterer. The depolarization defined in Eq. (14) arises because the orthogonal components of the scattered field undergo different attenuations and phase shifts. These differences are referred to as the differential attenuation and differential phase shift. An alternative definition related to depolarization is the cross-polarization discrimination, which is simply the reciprocal of the depolarization. In general, the depolarization increases as the particulate size and eccentricity increase. The depolarization also increases as the angle between the E field of the incident wave and the major axis of the particulate increases up

to approximately 45°, for which the depolarization passes through a maximum.

2.2.3. Types of Particulate. Rain is the most common particulate; drops range in size from a fraction of a millimeter to about 7 mm. Sleet and snow, which are considered quasisolid forms of water, are then treated. Because of their complexity of shape and composition, only limited theoretical work has been done on them, and because they are rare events in most locations, only limited experimental data have been obtained. Hail is frozen water, and does not occur very often. However, the losses due to hail can be calculated quite accurately, and are very small at millimeter waves because the complex dielectric constant of ice is small. Cloud, fog, and haze particulates are very similar in that they are all composed of very small water droplets suspended in air (clouds may also contain ice crystals) with diameters ranging from several microns (μ) up to about 100 μm. Therefore, through most of the millimeter wave region the Rayleigh approximation is valid for these particulates. Dust and sand particulates have size distributions comparable to that of clouds, but because their complex dielectric constants are low, their interaction with millimeter waves is very weak.

2.2.3.1. Rain. Rain is an extremely complex phenomenon, both meteorologically and electromagnetically. From a meteorologic standpoint it is generally nonuniform in shape, size, orientation, temperature, and distribution, thus making it very difficult to model. Electromagnetically, the absorption, scattering, and depolarization characteristics can be calculated only for very simple shapes and distributions. However, theoretical results do provide a qualitative understanding of the effects of rain on millimeter waves, and when they are combined with experimental data, empirical parameters can be derived and more quantitative results are possible. Let us first examine the absorption and scattering characteristics of a single spherical raindrop. From Eq. (15) we can calculate the total cross section Q_t, which is the sum of the absorption cross section Q_a and the scattering cross section Q_s. This cross section is a strong function of the drop diameter and its complex index of refraction. At millimeter wavelengths both the real and imaginary parts decrease with decreasing wavelength, and we shall see that this is one of the reasons why the cross section of a drop eventually starts to decrease at shorter wavelengths. In Fig. 3, the total cross section of a drop is plotted as a function of drop diameter for several wavelengths. When the drop is very small with respect to wavelength, the Rayleigh approximation is valid; it is seen from Eqs. (16) and (17) that the scattering and absorption cross sections Q_s, and Q_a are proportional to $(D/\lambda)^6$ and $(D/\lambda)^3$, respectively. Because the loss due to scattering is negligible compared to that due to absorption, the total cross section is proportional to the volume of the drop. As the drop increases in size, both the scattering and absorption cross sections continue to increase, with the scattering cross section increasing more rapidly. Finally, the total cross section begins to level off, and would eventually approach a value of twice the geometric cross section of the drop when it is very large with respect to wavelength [14]. Thus, as the drop becomes larger, the cross section, which is initially proportional to the drop volume, becomes proportional to the drop area. The dependence of the cross section on wavelength is more complicated than that of size, because both the relative drop size and the complex index of refraction are changing. In Fig. 4, the cross sections are plotted as a function of wavelength for a number of drop radii. The cross sections increase with decreasing wavelength, reach a peak, and then start to decrease very slightly for still smaller wavelengths. This behavior can be explained by considering that although the cross section increases as the drop becomes larger with respect to wavelength, the real and imaginary components of the index of refraction decrease as the wavelength becomes smaller, and this decrease eventually causes the total cross section to decrease.

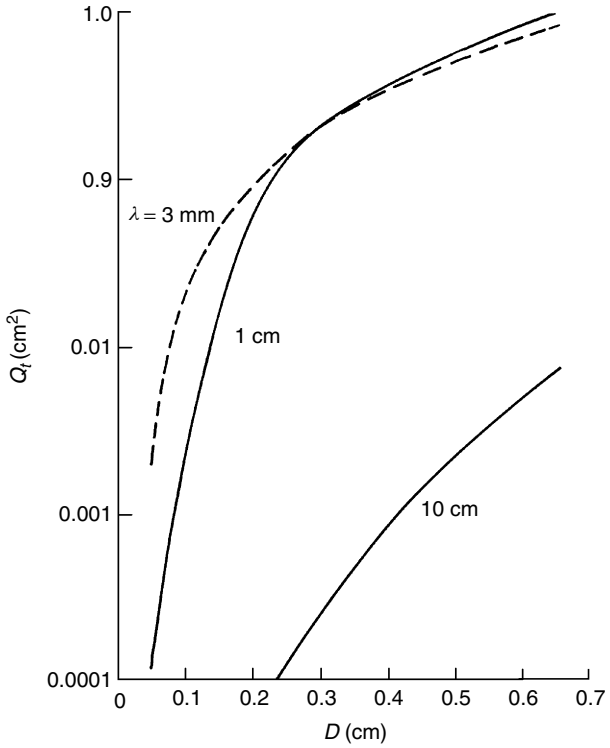


Figure 3. Total cross section Q_t of a raindrop as a function of drop diameter D.

We shall now consider the effect of the *shape* of raindrops on electromagnetic parameters. Whereas small drops tend to be spherical in shape, larger drops become oblate because of distortion due to air drag and are often modeled as oblate spheroids [15]. The cross section of an oblate drop is generally larger than that of a corresponding spherical drop having an equal volume of water [16]. The cross section is strongly dependent on the polarization of the wave, being larger when the polarization vector is aligned with the major axis of the spheroid and smaller when the polarization vector is aligned with the minor axis. We shall see that the most significant effect of a nonspherical drop is the depolarization it can produce.

We have reviewed the absorption and scattering characteristics of individual raindrops and found that they

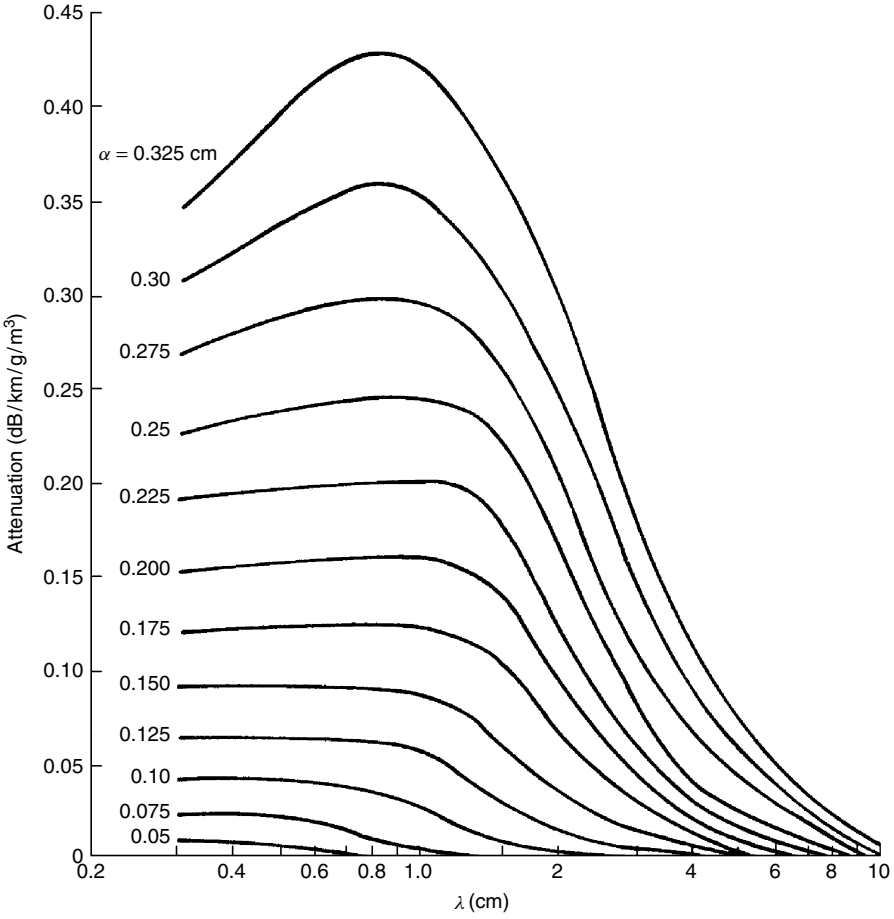


Figure 4. Theoretical values of attenuation by raindrops for various drop radii, expressed in decibels per kilometer per drop per cubic meter.

are a very complicated function of both the drop geometry and the index of refraction. Rain can be considered a collection of drops having diameters ranging from a fraction of a millimeter (mist) up to possibly 7 mm. To compute the attenuation of rain, the cross sections of the drops must be calculated and then summed. Because the characteristics of a precipitation system are controlled largely by the airflow, the net result is a collection of drops that is continually varying both spatially and temporally; it is thus very difficult to model. It has been found that the meteorologic parameter that is most easily measured and also most effectively characterizes rain is the rain rate. A number of investigators have shown that rain rate is correlated with drop size distribution; their results are summarized by Olsen et al. [17]. The attenuation can be expressed in the form

$$A = 0.4343 \int_0^\infty N(D) Q_t(D_t \lambda)\, dD \tag{19}$$

where $N(D)dD$ is the number of drops per cubic meter having diameters in the range dD, Q_t is the total cross section of each drop, and the attenuation is measured in decibels per kilometer. Rain is often assumed to have an exponential distribution of drop diameters, so that

$$N(D) = N_0\, e^{-\Lambda D}, \tag{20}$$

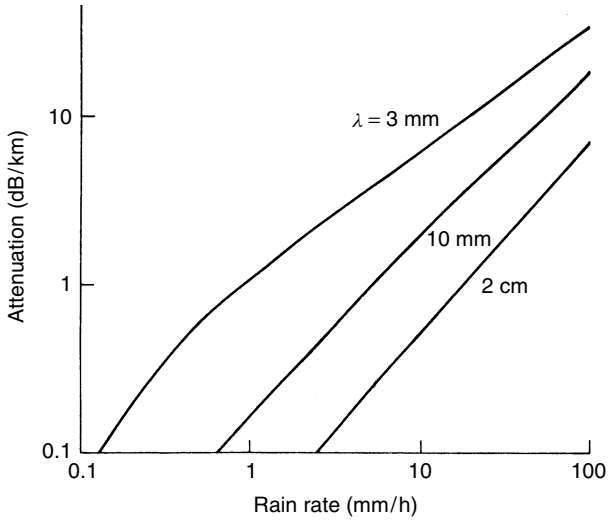


Figure 5. Rain attenuation as a function of rain rate.

where N_0 and Λ are empirical constants that are a function of the type of rain and more particularly the rain rate. Attenuations based on Medhurst's calculations [18] are plotted as a function of rain rate in Fig. 5. These curves can be approximated by

$$A = aR^b \tag{21}$$

where a and b are numerical constants that are a function of wavelength and type of rain, R is the rain rate in millimeters per hour, and the attenuation is measured in decibels per kilometer. Olsen et al. [17] calculated rain attenuation as a function of rain rate using the Mie formulation and then performed a logarithmic regression to obtain values for a and b. These values have been tabulated for frequencies from 1 to 1000 GHz, and although they are believed to provide a good approximation to the attenuation, it should be remembered that they are statistical and must be interpreted accordingly.

The depolarization characteristics of rain are very heavily dependent on the shape and orientation of the drops. Because light rain consists mostly of small drops and small drops tend to be spherical, depolarization effects are minimal. As the rain becomes heavier, the average drop size increases and the larger drops tend to become more oblate. The more oblate the drop, the larger the differential attenuation and phase between the orthogonal fields. However, it should be emphasized that these differentials do not in themselves produce depolarization; the incident field must also be tilted with respect to the axes of the drop. Because large oblates are easily canted by winds, their axes are seldom aligned with either horizontally or vertically polarized waves. Brussard [19] has shown that the canting angle of oblate raindrops is a function of the average drop diameter and vertical wind gradients. Typically the canting increases with drop size and levels off for drops on the order of a few millimeters in diameter. It naturally increases with wind speed and usually becomes smaller with increasing height. Because the differential attenuation is proportional to the total attenuation, it increases with rain rate. It also initially increases with shorter wavelengths, as does the total attenuation, but then reaches a peak and eventually starts to decrease at very short millimeter wavelengths, mostly because attenuation at very short wavelengths is produced primarily by the smaller drops and these tend to be spherical. The differential phase is affected mostly by the real part of the refractive index of water and decreases with shorter millimeter wavelengths. Thus, at millimeter wavelengths, differential attenuation is the dominant cause of depolarization

2.2.3.2. Sleet, Snow, and Hail. These very complex forms of precipitation have attenuation characteristics that vary markedly at millimeter wavelengths. We have seen that liquid water is a strong attenuator of millimeter waves; therefore sleet, which is a mixture of rain and snow, can also produce very high attenuations. In fact, these attenuations may exceed those of rain because nonspherical shapes have been shown to produce higher attenuations than equivalent spheres and sleet particulates are often very elongated [16]. The depolarization effects of sleet can be very strong if the flakes show a nonparallel preferential alignment with the E field. Wet snow has characteristics very similar to those of sleet. As the snow becomes drier, its composition approaches that of ice crystals, and because ice has a low imaginary index of refraction, the absorption is very small. Losses due to scattering are small at longer millimeter wavelengths but may become appreciable at shorter wavelengths if the flakes are large.

The effect of hail on millimeter waves is better understood than that of sleet or snow because there is less variability in its shape and composition. Because the imaginary part of the index of refraction of ice is about three orders of magnitude less than that of water, absorptive losses are negligible. The real part of the index of refraction is about one-fourth that of rain, so although scattering losses produced by hail are smaller than those of rain, they can be important, particularly at shorter millimeter wavelengths. If the hailstone is covered with even a very thin coat of water, its attenuation rises significantly and approaches that of a raindrop having an equivalent volume [20]. In summation, because it is extremely difficult to produce accurate models of sleet, snow, and hail, and because there are very few experimental attenuation data at millimeter wavelengths (or any other wavelengths), it is not possible to provide quantitative results. Although it is known that the absorption and scattering losses of sleet and wet snow are large and that the scattering losses of dry snow and hail are significant, there are presently no attenuation data for these particulates comparable to those for rain. However, from a practical standpoint, these particulates do not occur very often, and so their impact on millimeter wave systems is not considered critical.

2.2.3.3. Cloud, Fog, and Haze. Meteorologically, cloud, fog, and haze are very similar because they consist of small water droplets suspended in air. The droplets have diameters ranging from a fraction of a micron for fog and haze to over 100 μm for heavy fog and high-altitude clouds. Haze may be considered a light fog, and of course the principal difference between cloud and fog is that clouds exist at higher altitudes and often contain ice as well as water particulates.

Electromagnetically, cloud, fog, and haze can be treated identically. Because the droplets are very small with respect to wavelength, the Rayleigh approximation is valid, even at short millimeter wavelengths. Therefore, scattering losses are negligible and the attenuation, which is proportional to the volume of the drops, can be calculated from Eq. (17) and is seen to increase with decreasing wavelength. Because it is very difficult to measure fog or cloud water content, the attenuations produced by these particulates are not easily determined. Typical attenuations are plotted as a function of wavelength in Fig. 6 for several 'temperatures.

Fog is often characterized by visibility, which is much easier to measure than density; however, it should be emphasized that visibility is an optical parameter, a function of the scattering characteristics of the droplets, whereas the attenuation at millimeter wavelengths, as mentioned previously, is strictly a function of the fog density. Although there is a correlation between fog or cloud visibility and total liquid water content [21,22], this relationship must be used with caution because the correlation coefficient is strongly dependent on the type of fog. Fog is often divided into two types: *radiation fog*, which forms when the ground becomes cold at night and cools

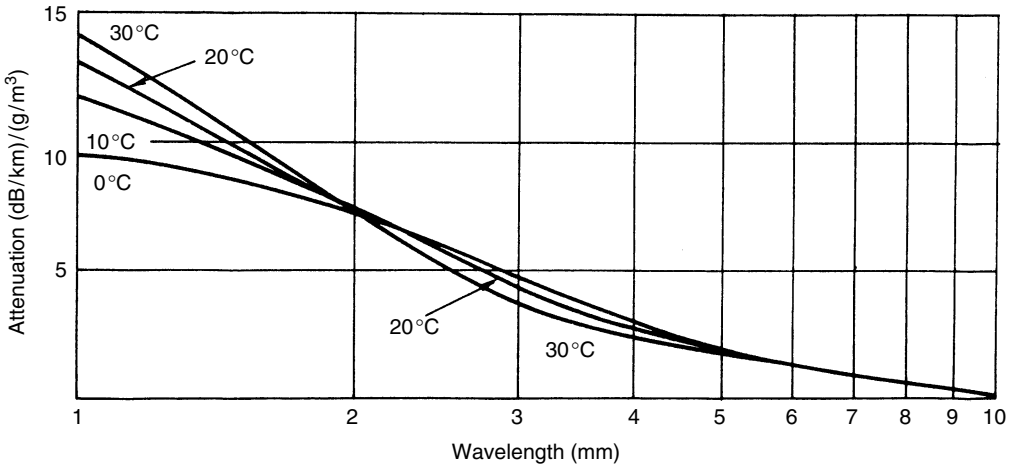


Figure 6. Attenuation of cloud and fog as a function of wavelength.

the adjacent air mass until it becomes supersaturated, and *advection fog*, which forms when warm, moist air moves across a cooler surface. The average drop size of an advection fog is usually larger than that of a radiation fog. Thus, if two fogs had the same liquid water densities but one consisted of relatively large drops, then that fog would have a much higher visibility. Another fog or cloud parameter not to be overlooked is temperature, because it has a strong influence on the complex index of refraction, particularly at longer millimeter wavelengths. Once again, as with snow and hail, the overall effects of cloud and fog are not as severe as those of rain.

2.2.3.4. Sand and Dust. Sand and dust are both fine-grained, quartz-type particles that have diameters of a fraction of a millimeter, densities on the order of 2600 kg/m^3, and a relative dielectric constant of about $2.5 - 0.025i$ [23]. The principal difference between sand and dust is that the average grain size of sand is larger. As a result, whereas larger windblown sand grains rise to a maximum height of only 2 m, heavy winds can carry fine dust particles to altitudes as high as 1 km. In general, the height to which the particles rise is proportional to the wind speed and inversely proportional to the particle size. Because sand and dust particulates are very small with respect to wavelength, even at millimeter wavelengths, scattering losses are negligible and the only losses are due to absorption. However, the imaginary part of the complex dielectric constant is very small; thus absorption losses can be considered minimal under naturally disturbed conditions. If a large amount of dirt or dust were to become suspended in the atmosphere as a result of an explosion, however, then significant attenuations might arise, particularly if the dust were moistened by the presence of water; these attenuations would last only for a duration of seconds.

3. TRANSMISSION PATHS

There are essentially two types of transmission paths: terrestrial and earth–space. We shall now take the results of Section 2 and apply them to systems that would use these propagation paths.

3.1. Terrestrial Line-of-Sight Paths

For terrestrial paths the atmospheric effects on propagation become more pronounced as the length of the path increases and the wavelength decreases. For short paths, attenuation is of principal concern; refraction and atmospheric multipath effects are unlikely, terrain multipath and diffraction problems arise only when transmitter and receiver are close to the surface, and depolarization and scintillations occur only under very extreme conditions of precipitation. For a clear atmosphere, gaseous absorption in the window regions is only a fraction of a decibel per kilometer at longer millimeter wavelengths but can become appreciable at shorter wavelengths. Sand and dust attenuations throughout the millimeter wave spectrum are well below 1 dB/km, except for conditions of a large cloud that may be produced by an explosion. Haze and fog attenuations increase with decreasing wavelength as shown in Fig. 6, and although they become appreciable, they are generally lower than water vapor absorption at very short wavelengths. Attenuation due to rain, sleet, and wet snow can be significant even for very short paths; this attenuation is lowest at longer wavelengths and gradually increases with decreasing wavelength, reaching a maximum at a wavelength of a few millimeters and then leveling off at still shorter wavelengths. As the path becomes longer, attenuation effects become more severe; in addition, all the other propagation effects mentioned in the previous paragraph are more likely to occur. The use of millimeter waves for applications requiring long terrestrial paths appears unlikely at this time because the attenuation would be prohibitive, except perhaps for a region having a relatively dry climate.

3.1.1. Attenuation. For many applications, particularly communications, it is important to be able to estimate the percentage of time that the path attenuation exceeds a certain value. To accomplish this one must first examine

the climate of the region of interest. If the absolute humidity is known, the gaseous absorption can be estimated from the expression

$$A = a + b\rho_0 - cT_0 \tag{22}$$

The coefficients are plotted in Fig. 7; ρ_0 is in grams per cubic meter and T_0 in degrees Celsius. This calculation is not very accurate at very short wavelengths, where there is a lack of agreement between the theoretical and experimental values of water vapor absorption. As mentioned in the previous paragraph, sand, dust, haze, and fog attenuations are low at longer wavelengths and generally small compared to water vapor absorption at very short wavelengths; only at wavelengths between about 2 and 3 mm is fog attenuation generally important.

Rain attenuation is by far the most serious propagation problem. Rain attenuation statistics are available only for some locations. A procedure for estimating rain attenuation for different climates has been outlined by Crane [24]. A global model representing typical rain climates throughout the world was developed, based on rain data provided by the World Meteorological Organization, and from this model the percentage of time that the point rain rate exceeds a particular value can be estimated. However, rain ordinarily consists of cells of different sizes and seldom is homogeneous in the horizontal plane. Thus, it is necessary to derive a path average rain rate from the point values. By pooling worldwide rain statistics it was possible to obtain an empirical relationship between point and path average rain rates; this has been done so far for distances up to 22.5 km. For low rain rates the rain is usually widespread; widespread rain, however, may contain convective cells having a high rain rate, so on an average the rain rate along the total path will be higher than that at a point. For high rain rates the rain tends to be localized, so the average rain rate for the total path would ordinarily be lower than that at a point. As expected, the correction factor is heavily dependent on the pathlength.

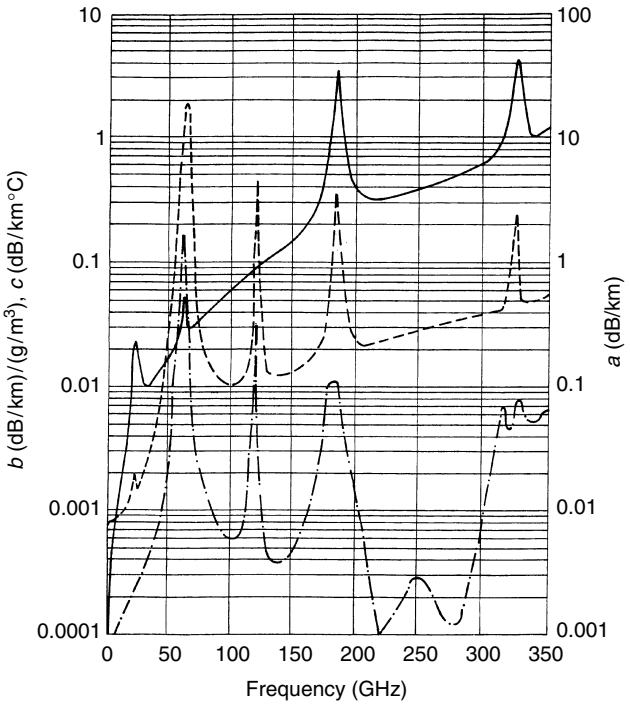


Figure 7. Coefficients for computing specific attenuation:- - - , a coefficient;———, b coefficient; —·—·, c coefficient.

In Section 2.2.3.1 the aR^b relationship for attenuation as a function of rain rate was introduced. By using the path average rain rate concept of the previous paragraph, it is possible to derive a correction term for the aR^b expression. From Crane [24], we have

$$A(R_p, D) = aR_p^b \left(\frac{e^{\beta d} - 1}{\mu\beta} - \frac{b^\beta e^{c\beta d}}{c\beta} + \frac{b^\beta e^{c\beta D}}{c\beta} \right) \qquad d \le D \le 22.5 \text{ km} \tag{23}$$

$$= aR_p^\beta \left(\frac{e^{u\beta D} - 1}{u\beta} \right) \qquad 0 < D \le d \tag{24}$$

where R_p is in millimeters per hour, D in kilometers, the specific attenuation $A(R_p, D)$ in decibels per kilometer, $\beta = 2\pi/\lambda$ and

$$u = \frac{\ln(be^{cd})}{d}, \qquad b = 2.3R_p^{-0.17}$$

$$c = 0.026 - 0.03 \ln R_p, \qquad d = 3.8 - 0.6 \ln R_p$$

For $D > 22.5$ km, the probability of occurrence P is replaced by a modified probability of occurrence

$$P' = \left(\frac{22.5}{D} \right) P \tag{25}$$

For example, suppose that for a particular climatic region the rain rate exceeds 28 mm/h 0.01% of the time and 41 mm/h 0.005% of the time ($R_{0.01} = 28$ mm/h, $R_{0.005} = 41$ mm/h). The attenuation along a 22.5-km path that would be exceeded 0.01% of the time can be calculated directly from Eq. (23) or (24). To determine the attenuation that would be exceeded 0.01% of the time over a 45-km path, a new percentage of time $P' = (22.5/45)P = 0.005\%$ would be used, with a corresponding rain rate of $R_{0.005} =$ 41 mm/h. Now the attenuation that would be exceeded 0.01% of the time for a 45-km path would be based on a rain rate of $R_{0.005} = 41$ mm/h for a 22.5-km path.

To improve the reliability of a terrestrial link, path diversity can be utilized. As mentioned previously, the heavy rain that has a severe impact on terrestrial link performance tends to be localized. By using redundant terminals, the probability of having a path free from heavy rain is increased. Ideally, the optimum separation of transmitter or receiver terminals (or both) is that for which the rain rates (and corresponding attenuations) for the pair of terminals are uncorrelated. This separation is a function of the climate and rain rate. Blomquist and Norbury [25] have studied diversity improvement for a number of paths with lengths of about 3–13 km and terminal separations of 4–12 km. On the basis of very limited data, they found that the diversity improvement increased as the terminal separation was increased from

4 to ~8 km; there was, however, no additional improvement as the separation was increased further. Because only limited statistical data on diversity advantage are presently available, it is not possible to determine optimal terminal spacing for most locations. However, there is sufficient evidence to indicate that a diversity mode of operation can significantly improve the performance of line-of-sight links.

3.1.2. Terrain Scatter and Diffraction. If the terminals of a line-of-sight path are close to the surface, then propagation losses due to multipath and diffraction are possible. These mechanisms were described in Sections 2.1.3.1 and 2.1.4, respectively, and it was seen that the multipath or diffracted signal can interfere with the direct signal; the net effect is a resultant signal that may vary in amplitude from zero intensity to twice the intensity of the direct signal. It should be emphasized that the interference is strongest when the multipath or diffracted signal is coherent. For multipath this occurs when the terrain is smooth with respect to wavelength; for diffraction it occurs from a prominent obstacle or a set of obstacles that happen to add constructively (or destructively). Regarding terrain multipath, even though most terrains have surface irregularities much larger than 1 mm, we see from Eqs. (8) and (9) that as the grazing angle becomes small, the surface becomes electromagnetically smooth (the reflection coefficient approaches unity) and large specular signals are possible. These signals interfere with the direct signal and can significantly degrade the performance of a line-of-sight system. Interference effects produced by a diffracted signal should not be as large as those produced by multipath, because the obstacles that would diffract the wave are not likely to produce a coherent signal. As mentioned previously, if the surface irregularities of a prominent obstacle are rough with respect to wavelength, there are many uncorrelated, diffracted rays and the resultant signal consists of a diffuse signal superimposed on a weak specular signal; if the surface is very rough, the specular component will disappear. Also, it is seen from Eq. (13) that the direct path must be within meters of the top of the obstacle for a diffracted signal to appear. However, diffracted signals are certainly possible at millimeter wavelengths under the "right" conditions.

3.1.3. Depolarization. As mentioned in Sections 2.1.5 and 2.2.2, depolarized signals may arise from either multipath or precipitation. A multipath ray obliquely incident on a paraboloidal receiving antenna can produce a cross-polarized component. Oblate raindrops canted with respect to the plane of the polarization vector of the incident wave also produce a cross-polarized component. The net effect is that the resultant signal is depolarized and the system performance compromised. Vander Vorst [26] has summarized the effects of depolarization on line-of-sight links. Often, depolarization produced by multipath may have a more severe effect on link performance than that produced by rain. The influence of rain caused only a very small degradation of the performance of a dual-polarized system with respect to a single-polarization system, whereas the same was not true for multipath.

3.1.4. Refraction and Atmospheric Multipath. Under normal atmospheric conditions an electromagnetic wave is bent toward the earth's surface. The amount of bending is proportional to the length of the path, so for long paths refractive bending corrections may be required, as discussed in Section 2.1.1.3. Under abnormal atmospheric conditions — for example, those producing a sharp negative gradient in the refractivity as a function of height — the wave may become trapped (ducting) or a multipath signal may be produced by the "layer" arising from the refractivity structure. Although refraction and multipath effects are possible in principle, they are not considered important as far as millimeter wave line-of-sight links are concerned, because, as mentioned previously, attenuation effects will normally prohibit the use of very long paths.

3.2. Earth–Space Paths

For earth–space paths, propagation effects become more severe with decreasing wavelength [27–29]. For elevation angles above about 6°, attenuation and emission from atmospheric gases and precipitation are of principal concern. In addition, backscatter and depolarization resulting from precipitation may also cause problems. For low elevation angles, all the problems associated with long terrestrial paths are present. The determination of propagation effects for slant paths is generally more difficult than for terrestrial paths. The modeling of a slant path under conditions of cloud and precipitation is particularly complicated, because the structure of the particulates is varying in both time and space. Experimentally, it is very expensive to place a millimeter wave beacon on a satellite or aircraft, so other means of obtaining attenuation information are sometimes used.

3.2.1. Attenuation and Emission. Atmospheric attenuation and emission are the most serious propagation problems for earth–space systems. Because the attenuation decreases the signal level and the emission (or effective noise temperature) sets a minimum noise level for the receiver, the only way to maintain the system signal-to-noise ratio is through increased transmitter power or a diversity mode of operation [30]. High powers are not readily available at millimeter wavelengths, and diversity systems are costly, so these options have their difficulties. For clear-sky conditions, the attenuation is a function of oxygen and water vapor density along the path. The vertical distributions of these gases are assumed to decrease exponentially with height and have scale heights of approximately 4 and 2 km, respectively. The zenith attenuation in decibels as a function of wavelength can be estimated from

$$A_{90^\circ} = \alpha + \beta\rho_0 - \xi T_0 \tag{26}$$

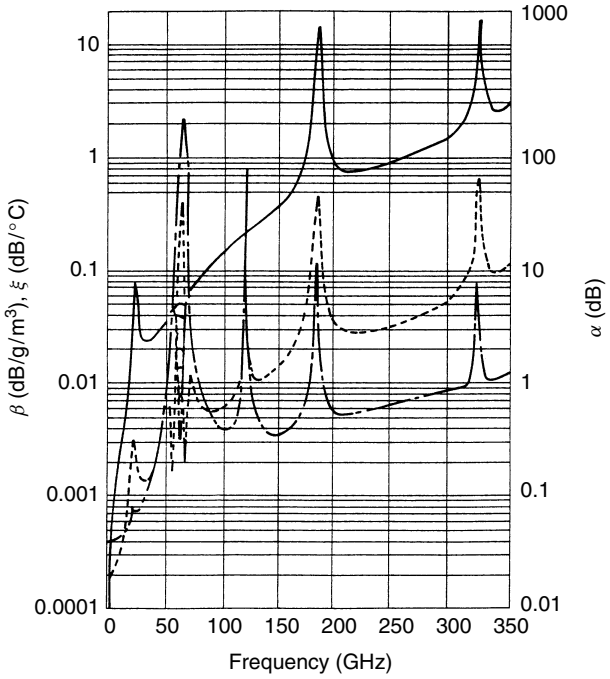


Figure 8. Coefficients for computing zenith attenuation: —·—·, α coefficient; ———, β coefficient; - - - , γ coefficient.

The coefficients α, β, and ξ are plotted in Fig. 8. The attenuation at elevation angles above ~6° can be calculated by multiplying the zenith attenuation by the cosecant of the elevation angle. For angles below 6° the attenuation is assumed to be proportional to the pathlength through the attenuating medium. This distance is given by Altshuler et al. [31] as

$$d(\theta) = [(a_e + h)^2 - a_e^2 \cos^2\theta]^{1/2} - a_e \sin\theta \qquad (27)$$

where θ is the elevation angle, a_e is $\frac{4}{3}$ the earth's radius ($a_e = 8500$ km), and h is the scale height of combined oxygen and water vapor gases (~3.2 km). Therefore

$$A(\theta) = \frac{A(90^\circ)d(\theta)}{h} \qquad (28)$$

The zenith attenuation is plotted as a function of frequency in Fig. 9 for a completely dry atmosphere and more typical atmospheres having surface absolute humidities of 3 and 10 g/m^3 [32]. For a dry atmosphere the total zenith attenuation in the window regions is only a fraction of a decibel. However, this attenuation increases very sharply as the atmosphere becomes moist, particularly below a wavelength of a few millimeters, at which losses on the order of tens of decibels are possible. A plot of the apparent sky temperature (emission) is shown in Fig. 10 as a function of frequency for a set of elevation angles from the horizon to zenith and for an atmosphere having a water vapor density of 7.5 g/m^3 [32]. The sky temperature is relatively low for higher-elevation angles and longer wavelengths and gradually increases for lower-elevation angles and

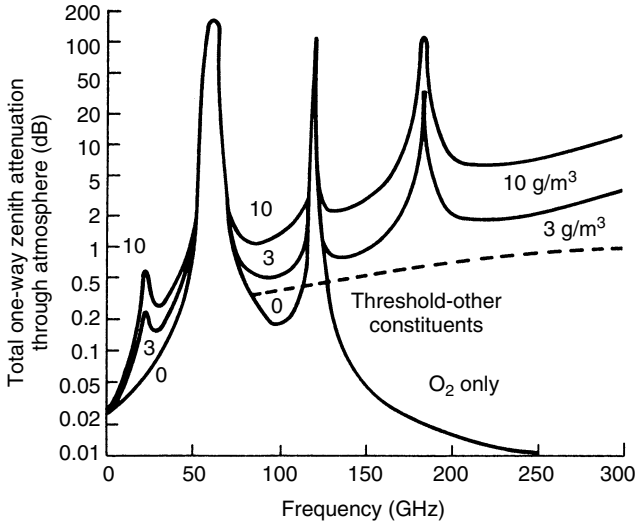


Figure 9. Total zenith attenuation through atmosphere as a function of frequency.

shorter wavelengths, approaching a terrain temperature of approximately 290 K.

For conditions of fog or cloud, the modeling of the atmosphere becomes increasingly difficult, particularly for slant paths close to the horizon. Fog and cloud models have been developed and the attenuation can be estimated using the information provided in Fig. 6. It must be emphasized that these attenuations are only approximate, because neither the true liquid water density of the cloud nor the extent of the cloud is accurately known. Because the cloud particulates are in the Rayleigh region and scattering losses are negligible, the corresponding brightness temperature (emission) can be calculated from Eq. (6). Several investigators have measured cloud attenuations at millimeter wavelengths. Altshuler et al. [31] have presented cloud attenuation statistics at frequencies of 15 and 35 GHz based on 440 sets of measured data. They characterized sky conditions as clear, mixed clouds, or heavy clouds. Average attenuations as a function of elevation angle are shown in Fig. 11. They also demonstrated a reasonable correlation between the slant path attenuation and surface absolute humidity. Typical slant path attenuations extrapolated to zenith were 0.1 and 0.36 dB at 15 and 35 GHz, respectively. Lo et al. [33] measured attenuations at wavelengths of 8.6 and 3.2 mm over a 6-month period, and obtained typical attenuations of 0.42 and 2.13 dB, respectively. Slobin [34] has estimated average-year statistics of cloud attenuation and emission down to a wavelength of 6 mm for various climatically distinct regions throughout the United States.

As was the case for terrestrial paths, rain, sleet, and wet snow present the most serious propagation limitations on earth–space millimeter wave systems. A number of investigators have derived models for predicting rain attenuation, and those results have been summarized by Ippolito [35] and Crane et al. [36–37]. All the techniques assume that the slant path attenuation can be estimated by modifying the attenuation for a terrestrial path by an

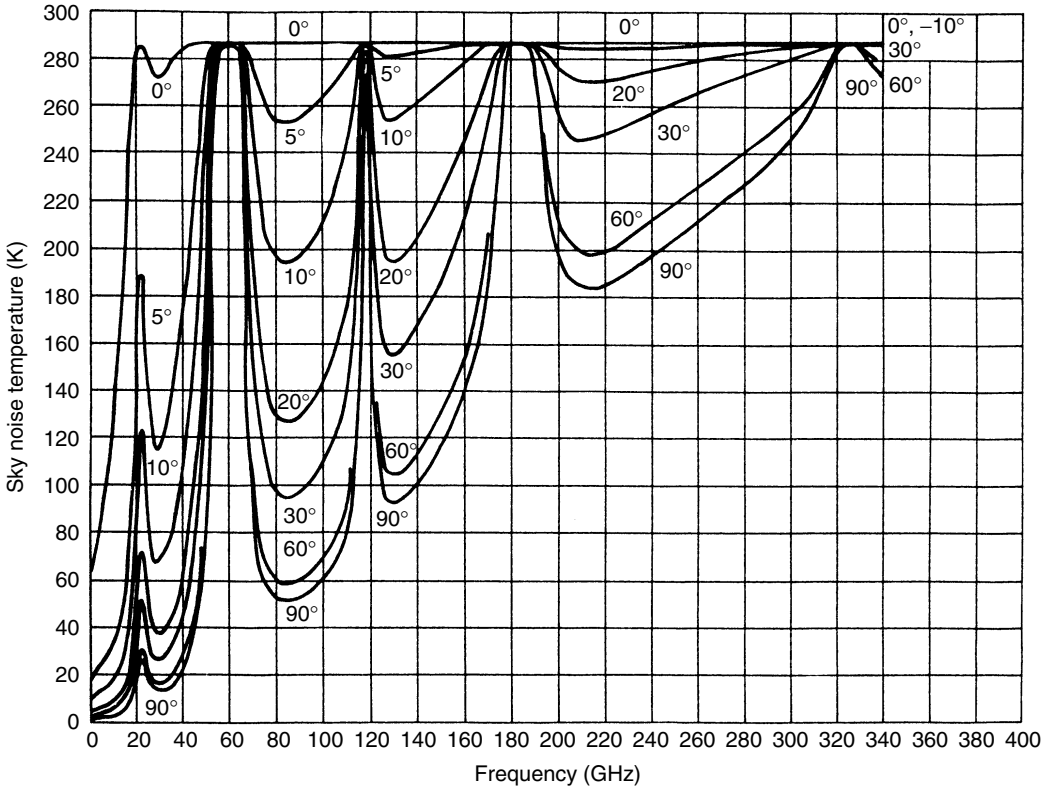


Figure 10. Sky noise temperature as a function of frequency for a water vapor density of 7.5 g/m^3.

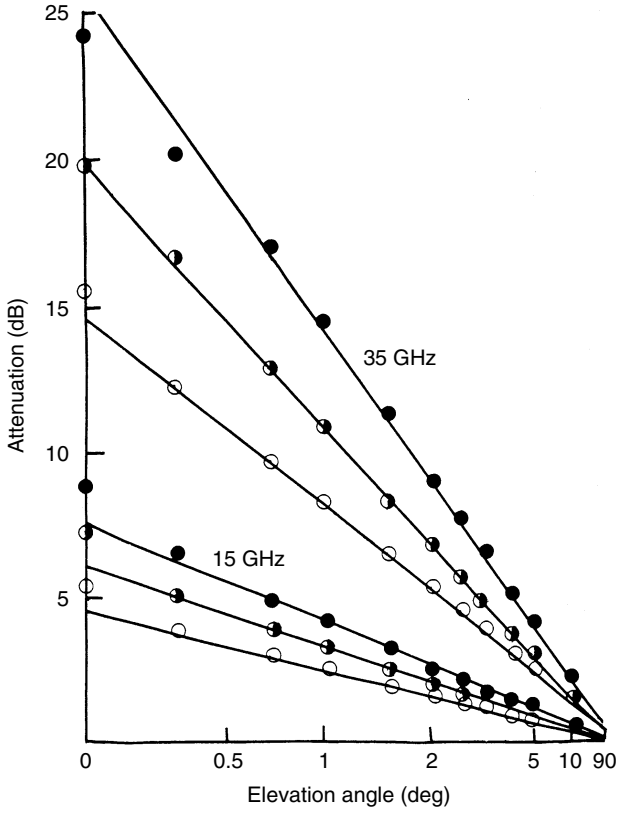


Figure 11. Typical cloud attenuations as a function of elevation angle: ☺ clear; ☻ cloudy; ◑ mixed.

effective pathlength parameter that is usually a function of the elevation angle and the type of rain. For example, the Crane model [24] for estimating rain attenuation for terrestrial paths, presented in Section 3.1.1, can be modified for slant path attenuations. It is assumed that the rain rate has a constant value between the station height h_0 and the height h of the 0 °C isotherm. Precipitation above h consists of ice particles, so the attenuation is considered negligible. Although the height of the 0 °C isotherm is variable, radar measurements have shown it to have a strong dependence on site latitude and rain rate, which in turn are linearly related to the logarithm of the probability of occurrence P. With h and h_0 known, the effective pathlength D through the rain can be calculated:

$$D = \frac{h - h_0}{\tan\theta} \qquad \theta \geq 10^\circ \tag{29}$$

$$= a_e \psi \qquad \theta < 10^\circ \tag{30}$$

where

$$\psi = \sin^{-1} \frac{\cos\theta}{h + a_e} \left\{ \begin{array}{l} [(h_0 + a_e)^2 \sin^2\theta + 2a_e(h - h_0) \\ +h^2 - h_0^2]^{1/2} - (h_0 + a_e)\sin\theta \end{array} \right\} \tag{31}$$

where a_e is the effective earth radius (8500 km) and θ is the elevation angle.

The surface-projected attenuation $A(R_p, D)$ is calculated from Eq. (23) or (24). Then, the value for the slant path A_s is estimated assuming a constant attenuation

below h by

$$A_s = \frac{LA(D)}{D} \tag{32}$$

$$L = \frac{D}{\cos\theta} \quad \theta \geq 10^\circ \tag{33}$$

$$= [(a_e + h_0)^2 + (a_e + h)^2 - 2(a_e + h_0)(a_e + h)\cos\psi]^{1/2}, \quad \theta < 10^\circ \tag{34}$$

Several methods can be used to measure slant path attenuations. The most straightforward, but also the most costly, is to place a millimeter wave beacon in space [38]. Measurements using satellite beacons have been made at wavelengths from approximately 30 to 10 mm and have been summarized by Ippolito [35] and Bauer [38]. A number of investigators have made attenuation measurements using the sun as a source [39–41]. When a radiometer is pointed at the sun, the noise power received consists of radiation from the sun and the atmosphere. The antenna temperature can be expressed as

$$T_a = T_a' e^{-\tau} + \int_0^\infty T(s)\gamma(s)\exp\left(-\int_0^\infty \gamma(s')\,ds'\right) ds \tag{35}$$

where T_a' is the effective antenna temperature of the sun with no intervening atmosphere (in degrees Kelvin), $T(s)$ is the atmospheric temperature, τ is the total attenuation (in nepers), $\gamma(s)$ is the absorption coefficient, and s is the distance from the antenna (ray path). In simpler terms

$$T_a = T_a' e^{-\gamma} + (1 - e^{-\gamma})T_m \tag{36}$$

where T_m is the atmospheric mean absorption temperature within the antenna beam. The attenuation γ appears in both terms on the right-hand side of Eq. (36). Because the second term is the emission, it can easily be canceled out by pointing the antenna beam toward and away from the sun. With the second term balanced off, Eq. (36) can be solved for γ, converted from nepers to decibels, and expressed as

$$A = 10\log\left(\frac{T_a'}{T_a}\right) \tag{37}$$

T_a' is determined from a set of antenna temperature measurements made under clear-sky conditions as a function of elevation angle; for these conditions the antenna temperature is proportional to the cosecant of the elevation angle, and it can be shown that T_a' is equal to the slope of the line $\log T_a$ versus $\csc\theta$ [41]. Because the attenuation also appears in the emission term in Eq. (36), it is possible to determine the attenuation from an emission measurement. In this method the antenna must be pointed away from the sun or the moon (millimeter wave radiation from all other natural sources is negligible), so the first term on the right-hand side of Eq. (36) can be considered zero and Eq. (36) reduces to Eq. (6). As before, the equation is then solved for the attenuation γ and expressed in decibels as

$$A = \frac{10\log T_m}{T_m - T_a} \tag{38}$$

Another technique for estimating rain attenuation at millimeter wavelengths is from a measurement of the reflectivity factor of the rain. Attenuation is derived from established relationships between these parameters [42,43]. McCormick [44] has compared attenuations derived from radar reflectivity with those obtained directly utilizing a beacon placed on an aircraft. Strickland [45] has measured slant path attenuations using radar, radiometers, and a satellite beacon simultaneously. When a single-wavelength radar is used, calibration errors and the uncertainty in the reflectivity–attenuation relationship, particularly for mixed-phase precipitation, limit the accuracy of this technique. Uncertainties in the reflectivity–attenuation relationship can, however, be reduced by using a dual-wavelength radar and measuring the differential attenuation.

In summation, there are four methods for measuring slant path attenuations. The most direct method is to place a source in space. This allows measurements to be made over a very wide dynamic range. An additional advantage is that the polarization and bandwidth limitations imposed by the atmosphere can also be measured. One disadvantage is cost; and, depending on the satellite orbit, measurements may be possible at only one elevation angle, which may or may not be a drawback. Total attenuation can be measured very accurately and economically using the sun as a source over dynamic ranges approaching 25–30 dB. A disadvantage is that measurements can be made only in the direction of the sun and during the day. Attenuation can easily be determined from an emission measurement on a continual basis and at any elevation angle. It must be emphasized that there are two major problems that limit the accuracy of this technique. The true value of T_m is not always known; if $T_m - T_a$ is large, this uncertainty is not serious, but if $T_m - T_a$ is small, a large error may arise. Also, the emission is related only to the absorption, whereas the attenuation includes losses due to scattering in addition to those due to absorption. Therefore, in cases for which the Rayleigh approximation is not valid and scattering losses are appreciable, errors will arise. Techniques for correcting for the additional losses due to scattering have been investigated by Zavody [46] and Ishimaru and Cheung [47]. For these reasons this method is not generally recommended for attenuations much above 10 dB. Attenuations determined from radar reflectivity measurements have the limitations in accuracy discussed previously, and in general this technique is not suitable for losses arising from very small particulates such as fog, cloud, or drizzle. This method does, however, have the advantage of measuring attenuation as a function of distance from the transmitter.

4. CONCLUSION

The interaction of millimeter waves with atmospheric gases and particulates has been examined. Precipitation in general and rain in particular limit the performance of longer millimeter wave systems. Systems operating at short millimeter wavelengths are significantly affected by

high water vapor absorption in addition to precipitation, so applications in this region of the spectrum will of necessity be limited to very short paths.

The use of millimeter waves has probably not progressed as rapidly as had been originally anticipated. For many years, the more optimistically inclined envisioned millimeter waves revolutionizing traditionally longer wavelength communications. When the discovery of the laser created a temporary lull in millimeter wave research, the more pessimistically inclined feared that millimeter waves had passed from infancy to obsolescence without having experienced a period of fruitfulness. Finally, there were realists who recognized that cost is a major consideration and that millimeter wave systems would reach the marketplace only when they could be shown to be competitive with systems that operate at longer or shorter wavelengths or to have unique properties such that needed applications could be realized only with millimeter waves. So far, history seems to be supporting the realists.

BIOGRAPHY

Edward E. Altshuler received a B.S. degree in physics from Northeastern University, Boston, Massachusetts, in 1953, an M.S. degree in physics from Tufts University, Medford, Massachusetts, in 1954, and the PhD. degree in applied physics from Harvard University, Cambridge, Massachusetts, in 1960. He joined Air Force Cambridge Research Labs (AFCRL), Hanscom Air Force Base (AFB), Massachusetts in 1960, but left in 1961 to become director of engineering at Gabriel Electronics, Millis, Massachusetts; he later returned to AFCRL in 1963 as chief of the propagation branch from 1963 to 1982. He was a lecturer in the Northeastern University Graduate School of Engineering from 1964 to 1991. He has served on the Air Force Scientific Advisory Board and was chairman of the NATO Research Study Group on millimeter wave propagation from 1974 to 1993. He was President of the Hanscom Chapter of Sigma Xi during 1989 through 1990. He received the IEEE Harry Diamond Memorial Award in 1997 and was awarded an IEEE Millennium Medal in 2000. He is a fellow of both the IEEE and AFRL. Dr. Altshuler has over 120 scientific publications, conference papers, and patents. He is currently conducting antenna research for the Air Force Research Laboratory at Hanscom AFB.

BIBLIOGRAPHY

1. E. E. Altshuler, New applications at millimeter wavelengths, *Microwave J.* **11**: 38–42 (1968).
2. E. K. Smith and S. Weintraub, The constants in the equation for atmospheric refractive index at radio frequencies, *Proc. IRE* **41**: 1035–1037 (1953).
3. B. R. Bean and E. J. Dutton, *Radio Meteorology*, Dover, New York, 1968.
4. E. E. Altshuler, Tropospheric range-error corrections for the global positioning system, *IEEE Trans. Antennas Propag.* **46**: 643–649 (1998).
5. G. K. Elgered, Tropospheric wet-path delay measurements, *IEEE Trans. Antennas Propag.* **30**: 502–505 (1982).
6. M. A. Gallop, Jr. and L. E. Telford, Use of atmospheric emission to estimate refractive errors in a non-horizontally stratified troposphere, *Radio Sci.* **11**: 935–945 (1975).
7. L. W. Schaper, Jr., D. H. Staelin, and J. W. Waters, The estimation of tropospheric electrical path length by microwave radiometry, *Proc. IEEE* **58**: 272–273 (1970).
8. S. C. Wu, Optimum frequencies of a passive microwave radiometer for tropospheric path-length correction, *IEEE Trans. Antennas Propag.* **27**: 233–239 (1979).
9. J. Goldhirsh, B. H. Musiani, and W J. Vogel, Cumulative fade distributions and frequency scaling techniques at 20 GHz from the advanced communications technology satellite and at 12 GHz from the digital satellite system, *Proc. IEEE* **85**: 910–916 (1997).
10. C. E. Mayer, B. E. Jaeger, R. K. Crane, and X. Wang, Ka-band scintillations: measurements and model predictions, *Proc. IEEE* **85**: 936–945 (1997).
11. F. S Marzano and C. Riva, Evidence of long-term correlation between clear-air attenuation and scintillation in microwave and millimeter-wave satellite links, *IEEE Trans. Antennas Propag.* **47**: 1749–1757 (1979).
12. J. W. Waters, *Methods of Experimental Physics*, 12B, Academic Press, New York, 1976, Chap. 23.
13. R. L. Olsen, Cross polarization during clear-air conditions on terrestrial links-a review, *Radio Sci.* **16**: 631–647 (1981).
14. H. C. Van de Hulst, *Light Scattering by Small Particles*, Wiley, New York, 1957.
15. H. R. Pruppacher and R. L. Pitter, A semi-empirical determination of the shape of cloud and rain drops, *J. Atmos. Sci.* **28**: 86–94 (1971).
16. D. Atlas, M. Kerker, and W. Hitschfeld, Scattering and attenuation by nonspherical atmospheric particles, *J. Atmos. Terr. Phys.* **3**: 108–119 (1953).
17. R. L. Olsen, D. V. Rogers, and D. E. Hodge, The aRb relation in the calculation of rain attenuation, *IEEE Trans. Antennas Propag.* **26**: 318–329 (1978).
18. R. G. Medhurst, Rainfall attenuation of centimeter waves: Comparison of theory and measurement, *IEEE Trans. Antennas Propag.* **13**: 550–563 (1965).
19. G. Brussaard, A meteorological model for rain-induced cross polarization, *IEEE Trans. Antennas Propag.* **24**: 5–11 (1976).
20. L. J. Battan, *Radar Observations of the Atmosphere*, Univ. Chicago Press, 1973.
21. R. G. Eldridge, Haze and fog aerosol distributions, *J. Atmos. Sci.* **23**: 605–613 (1966).
22. C. Platt, Transmission of submillimeter waves through water clouds and fogs, *J. Atmos. Sci.* **27**: 421–425 (1970).
23. T. S. Chu, Effects of sandstorms on microwave propagation, *Bell Syst. Tech. J.* **58**: 549–555 (1979).
24. R. K Crane, Prediction of attenuation by rain, *IEEE Trans. Commun.* **28**: 1717–1733 (1980).
25. A. Blomquist and J. R. Norbury, Attenuation due to rain or series, parallel and convergent terrestrial paths, *Alta Freq.* **66**: 185–190 (1979).
26. A. VanderVorst, Cross polarization on a terrestrial path, *Alta Freq.* **48**: 201–209 (1979).

27. D. V Rogers, L. J. Ippolito, Jr., and F. Davarian, System requirements for Ka-band earth-satellite propagation data, *Proc. IEEE* **85**: 810–820 (1997).

28. Y. Karasawa and Y. Maekawa, Ka-band earth-space propagation research in Japan, *Proc. IEEE* **85**: 821–842 (1997).

29. R. Arbesser-Rastburg and A. Paraboni, European research on Ka-band slant path propagation, *Proc. IEEE* **85**: 843–852 (1997).

30. H. Helmken et al., A three-site comparison of fade-duration measurements, *Proc. IEEE* **85**: 917–925 (1997).

31. E. E. Altshuler, M. A. Gallop, and L. E. Telford, Atmospheric attenuation statistics at 15 and 35 GHz for very low elevation angles, *Radio Sci.* **13**: 839–852 (1978).

32. E. K. Smith, Centimeter and millimeter wave attenuation and brightness temperature due to atmospheric oxygen and water vapor, *Radio Sci.* **17**: 1455–1464 (1982).

33. L. Lo, B. M. Fanning, and A. W. Straiton, Attenuation of 8.6 and 3.2 mm radio waves by clouds, *IEEE Trans. Antennas Propag.* **23**: 782–786 (1975).

34. S. D. Slobin, Microwave noise temperature and attenuation of clouds: Statistics of these effects at various sites in the United States, Alaska and Hawaii, *Radio Sci.* **17**: 1443–1454 (1982).

35. L. J. Ippolito, Radio propagation for space communications systems, *Proc. IEEE* **69**: 697–727 (1981).

36. R. K. Crane and A. W Dissanayake, ACTS propagation experiment: Attenuation distribution observations and prediction model comparisons, *Proc. IEEE* **85**: 879–892 (1997).

37. R. K. Crane and P. C. Robinson, ACTS propagation experiment: Rain-rate distribution observations and prediction model comparisons, *Proc. IEEE* **85**: 946–958 (1997).

38. R. Bauer, Ka-band propagation measurements: An opportunity with the advanced communications technology satellite (ACTS), *Proc. IEEE* **85**: 853–862 (1997).

39. E. E. Altshuler and L. E. Telford, Frequency dependence of slant path rain attenuations at 15 and 35 GHz, *Radio Sci.* **15**: 781–796 (1980).

40. R. W. Wilson, Suntracker measurements of attenuation by rain at 16 and 30 GHz, *Bell Syst. Tech. J.* **48**: 1383–1404 (1969).

41. K. N. Wulfsberg, Atmospheric attenuation at millimeter wavelengths, *Radio Sci.* **2**: 319–324 (1967).

42. J. Goldhirsh, A review on the application of non-attenuating frequency radars for estimating rain attenuation and space-diversity performance, *IEEE Trans. Geosci. Electron.* **17**: 218–239 (1979).

43. J. D. Beaver and V. N. Bringi, The application of S-band polarimetric radar measurements to Ka-band attenuation prediction, *Proc. IEEE* **85**: 893–909 (1997).

44. K. S. McCormick, A comparison of precipitation attenuation and radar backscatter along earth-space paths, *IEEE Trans. Antennas Propag.* **20**: 747–755 (1972).

45. J. I. Strickland, The measurement of slant path attenuation using radar, radiometers and a satellite beacon, *J. Rech. Atmos.* **VIII**: 347–358 (1974).

46. A. M. Zavody, Effect of scattering by rain or radiometer measurements at millimeter wavelengths, *Proc. IEE* **121**: 257–263 (1974).

47. A. Ishimaru and R. L. T. Cheung, Multiple-scattering effect on radiometric determination of rain attenuation at millimeter wavelengths, *Radio Sci.* **15**: 507–516 (1980).

MIMO COMMUNICATION SYSTEMS

ROHIT U. NABAR
AROGYASWAMI J. PAULRAJ
Stanford University
Stanford, California

1. INTRODUCTION

Successful deployment of wireless networks presents a number of challenges. These include limited availability of the radiofrequency spectrum and a complex time-varying wireless environment (fading and multipath). Meeting the increasing demand for higher data rates, better quality of service (QoS), fewer dropped calls, higher network capacity, and user coverage calls for innovative techniques that improve spectral efficiency and link reliability. The use of multiple antennas at both receiver and transmitter in a wireless system is an emerging technique that promises significant improvements in these measures. This technology, popularly known as multiple-input/multiple-output (MIMO) wireless technology, offers a variety of (often competing) leverages that, if exploited correctly, can significantly improve network performance. These leverages include *array gain, diversity gain, multiplexing gain*, and *interference reduction*.

Figure 1 shows a typical MIMO system with M_T transmit antennas and M_R receive antennas. The space–time modem at the transmitter (Tx) encodes and modulates the information bits to be conveyed to the receiver. Additionally, it maps the signals to be transmitted across space (M_T transmit antennas) and time. The space–time modem at the receiver (Rx) processes the signals received on each of the M_R receive antennas, and in accordance with the transmitter's signaling strategy, demodulates and decodes the received signal. Signaling strategies are designed on the basis of the transmitter's knowledge of the wireless channel (we assume channel knowledge at the receiver) and link requirements (data rate, error rate, etc.) and exploit one

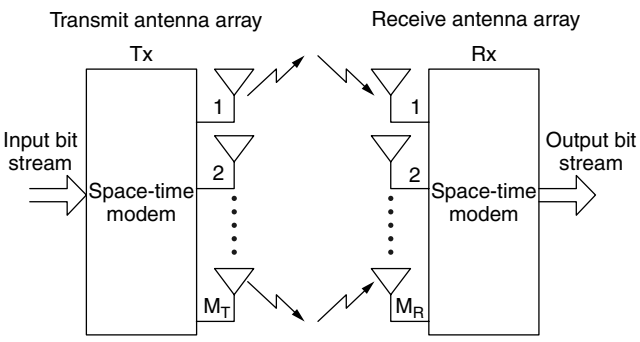


Figure 1. Schematic of a MIMO communication system.

or more of the four leverages of MIMO systems. In the following section, we describe these key leverages.

2. LEVERAGES OF MIMO TECHNOLOGY

To explore the leverages of MIMO technology we focus on various subsets of a MIMO link including MISO (multiple input/single output), SIMO (single input/multiple output) and SISO (single input/single output). Leverages such as array gain, diversity gain, and interference reduction that are available in MIMO systems are also offered by SIMO and MISO systems. Multiplexing gain, however, can only be exploited in MIMO systems.

2.1. Array Gain

Consider a SIMO system with one transmit antenna and two receive antennas as shown in Fig. 2. The two receive antennas see different versions, s_1 and s_2, of the same transmitted signal, s. The signals s_1 and s_2 have different amplitudes and phases as determined by the propagation conditions. If the channel is known to the receiver, appropriate signal processing techniques can be applied to combine the signals s_1 and s_2 coherently so that the resultant power of the signal at the receiver is enhanced, leading to an improvement in signal quality. More specifically, the signal-to-noise ratio [ratio of signal power to noise power (SNR)] at the output is equal to the sum of the SNR on the individual links. This result can be extended to systems with one transmit antenna and more than two receive antennas. The average increase in signal power at the receiver in such systems is defined as array gain and is proportional to the number of receive antennas. Array gain can also be exploited in systems with multiple antennas at the transmitter (MISO or MIMO systems). Extracting the maximum possible array gain in such systems requires channel knowledge at the transmitter, so that the signals may be optimally processed before transmission. Analogous to the SIMO case, the array gain in MISO systems is proportional to the number of transmit antennas. The array gain in MIMO systems depends on the number of transmit and receive antennas and is a function of the dominant singular value of the channel.

2.2. Diversity Gain

Signal power in a wireless channel fluctuates (or fades) with time–frequency–space. When the signal power drops dramatically, the channel is said to be in a fade. Diversity is used in wireless systems to combat fading. The basic principle behind diversity is to provide the receiver with several looks at the transmitted signal over independently fading links (or diversity branches). As the number of diversity branches increases, the probability that at any instant of time one or more branch is not in a fade increases. Thus diversity helps stabilize a wireless link.

Diversity is available in SISO links in the form of time or frequency diversity. The use of time or frequency diversity in SISO systems often incurs a penalty in data rate due to the utilization of time or bandwidth to introduce redundancy. The introduction of multiple antennas at the transmitter and/or receiver provides spatial diversity, the use of which does not incur a penalty in data rate while adding the array gain advantage discussed earlier. In this article we are concerned with this form of diversity. To utilize spatial diversity we must transmit and receive from antennas that are spaced by more than the coherence distance, which is the minimum spatial separation between antennas at the receiver (and/or transmitter) that ensures that the received signals (or their components) experience independent fading. In a rich scattering environment the coherence distance is approximately equal to half the wavelength ($\lambda/2$) [1] of the transmitted signal. There are two forms of spatial diversity: receive and transmit diversity.

Receive diversity applies to systems with multiple antennas only at the receiver (SIMO systems) [2]. Figure 3 illustrates a system with receive diversity. Signal s is transmitted from a single antenna at the transmitter. The two receive antennas see independently faded versions, s_1 and s_2, of the transmitted signal, s. The receiver combines these signals using appropriate signal processing techniques so that the resultant signal exhibits greatly reduced amplitude variability (fading) as compared to either s_1 or s_2. The amplitude variability can be further reduced by adding more antennas to the receiver. The diversity in a system is characterized by the number of independently fading diversity branches, also known as the diversity order. The diversity order of the system in Fig. 3 is two and in general is equal to the number of receive antennas, M_R, in a SIMO system.

Transmit diversity is applicable when multiple antennas are used at the transmitter and has become an active

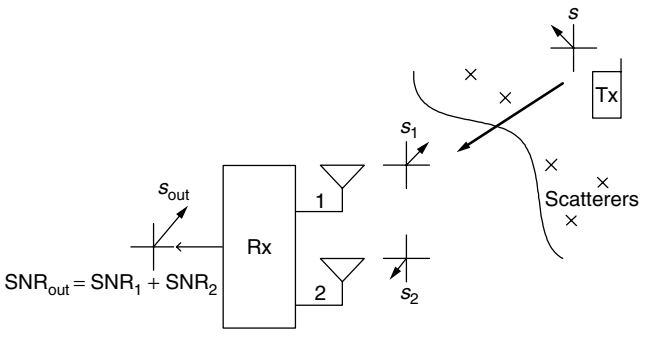


Figure 2. Array gain.

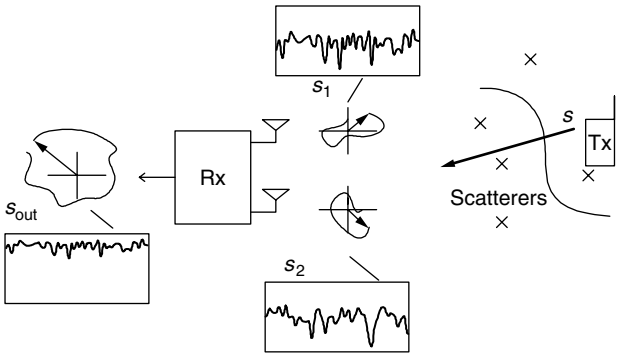


Figure 3. Receive diversity.

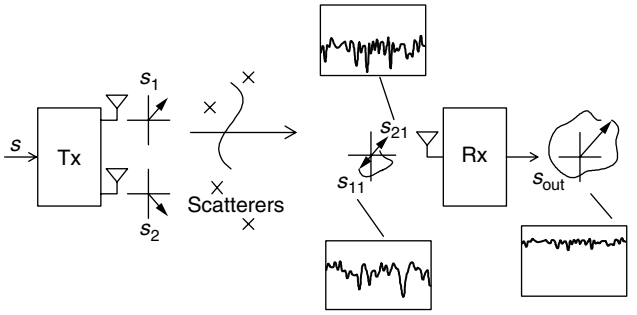


Figure 4. Transmit diversity.

area for research [3–5]. Extracting diversity in such systems does not necessarily require channel knowledge at the transmitter. However, suitable design of the transmitted signal is required to extract diversity. Space–time coding [6,7] is a powerful transmit diversity technique that relies on coding across space (transmit antennas) and time to extract diversity. Figure 4 shows a generic transmit diversity scheme for a system with two transmit antennas and one receive antenna. At the transmitter, signals s_1 and s_2 are derived from the original signal to be transmitted, s, such that the signal s can be recovered from either of the received signals s_{11} or s_{21}. The receiver combines the received signals in such a manner that the resultant output exhibits reduced fading when compared to s_{11} or s_{21}. The diversity order of this system is two and in general is equal to the number of transmit antennas, M_T, in a MISO system.

Utilization of diversity in MIMO systems requires a combination of receive and transmit diversity described above. A MIMO system can be decomposed into $M_T \times M_R$ SISO links. If the signals transmitted over each of these links experience independent fading, then the diversity order of the system is given by $M_T \times M_R$. Thus the diversity order in a MIMO system scales linearly with the product of the number of receive and transmit antennas.

2.3. Multiplexing Gain

MIMO systems offer a capacity (data rate) enhancing leverage not available in SIMO or MISO systems. We refer to this leverage as multiplexing gain which can be realized through a technique known as *spatial multiplexing* [8,9]. Figure 5 shows the basic principle of spatial multiplexing for a system with two transmit and two receive antennas. The symbol stream to be transmitted is split into two half-rate substreams and modulated to form the signals s_1 and s_2 that are transmitted simultaneously from separate antennas. Under favorable channel conditions, the spatial signatures of these signals [denoted by $[s_{11} \quad s_{12}]^T$ (the superscript ^{T} represents matrix transpose) and $[s_{21} \quad s_{22}]^T$] induced at the receive antennas are well separated (ideally orthogonal). The receiver can then extract the two substreams, s_1 and s_2, which it combines to give the original symbol stream, s.

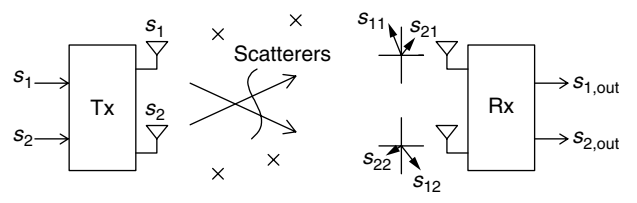


Figure 5. Spatial multiplexing.

2.4. Interference Reduction

Cochannel interference arises due to the reuse of frequency spectrum in wireless networks and adds to the overall noise in the system and deteriorates performance. Figure 6 illustrates the general principle of interference reduction for a receiver with two antennas. Typically, the desired signal (s) and the interference (i) arrive at the receiver with well separated spatial signatures — $[s_1 \quad s_2]^T$ and $[i_1 \quad i_2]^T$, respectively. The receiver can exploit the difference in signatures to reduce the interference, thereby enhancing the signal to interference ratio [ratio of signal power to interference power (SIR)]. Interference reduction requires knowledge of the desired signal's channel. Complete knowledge of the interfering signal's channel is not necessary. Interference reduction can also be implemented at the transmitter, where the goal is to enhance the signal power at the intended receiver and minimize the interference energy sent toward the cochannel users. Interference reduction allows the use of aggressive reuse factors and improves network capacity.

Having discussed the key advantages of MIMO technology we note that it may not be possible to exploit all the leverages simultaneously in a MIMO system. This is because some of the leverages and their methods of realization may be mutually conflicting. The optimal MIMO signaling strategy is a function of the wireless channel and network requirements. Exploiting the benefits of MIMO technology requires a good understanding of the MIMO channel. In the following section we introduce a simple MIMO channel model for an interference free environment.

3. MIMO CHANNEL MODEL

Consider a MIMO system with M_T transmit antennas and M_R receive antennas as shown in Fig. 7. For simplicity

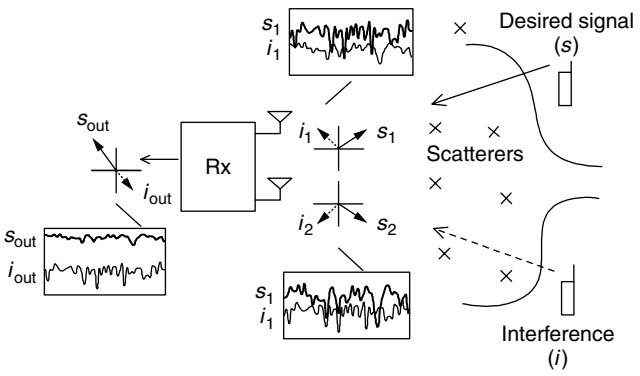


Figure 6. Interference reduction.

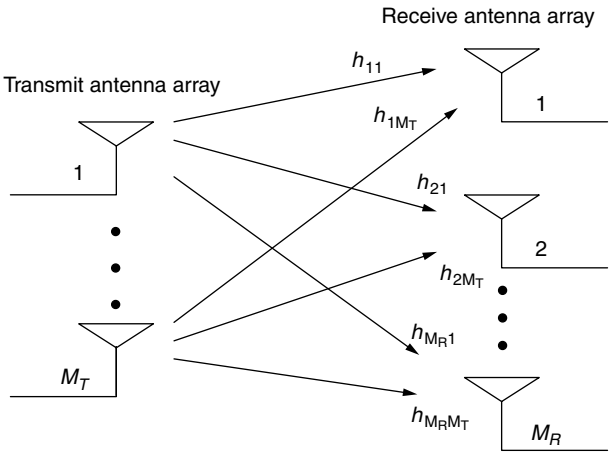


Figure 7. Flat fading MIMO channel model.

we consider only frequency flat fading; thus, the fading is not frequency-selective. When a continuous-wave (CW) probing signal, s, is launched from the jth transmit antenna, each of the M_R receive antennas see a complex weighted version of the transmitted signal. We denote the signal received at the ith receive antenna by $h_{ij}s$, where h_{ij} is the channel response between the jth transmit antenna and the ith receive antenna. The vector $[h_{1j}\, h_{2j} \cdots h_{M_Rj}]^T$ is the signature induced by the jth transmit antenna across the receive antenna array. It is convenient to denote the MIMO channel ($\mathbf{H}$) in matrix notation as shown below:

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1M_T} \\ h_{21} & h_{22} & \cdots & h_{2M_T} \\ \vdots & \vdots & \ddots & \vdots \\ h_{M_R1} & h_{M_R2} & \cdots & h_{M_RM_T} \end{bmatrix} \tag{1}$$

The channel matrix $\mathbf{H}$ defines the input–output relation of the MIMO system and is also known as the channel transfer function. If a signal vector $\mathbf{x} = [x_1\, x_2 \cdots x_{M_T}]^T$ is launched from the transmit antenna array (x_j is launched from the jth transmit antenna) then the signal received at the receive antenna array, $\mathbf{y} = [y_1 y_2 \cdots y_{M_R}]^T$ is given by

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n} \tag{2}$$

where $\mathbf{n}$ is the $M_R \times 1$ noise vector consisting of independent complex Gaussian distributed elements with zero mean and variance σ_n^2 (white noise). Note that the discussion above pertains to a snapshot of the channel at a particular frequency and a specific instant of time. Channels with large delay spreads show greater variability of $\mathbf{H}$ with frequency. Likewise, channels with large Doppler spreads show greater variability of $\mathbf{H}$ with time. In a scattering environment with sufficient antenna separation at the transmitter and receiver, the elements of the channel matrix $\mathbf{H}$ can be assumed to be independent, zero-mean, complex Gaussian random variables (Rayleigh fading) with equal variances. This model is popularly referred to as the i.i.d. fading MIMO channel model.

The choice of MIMO signaling strategies that optimize performance depends on knowledge of the channel at the receiver and/or transmitter. Channel knowledge at the receiver is a common assumption. Knowledge of the channel can be maintained at the receiver via training and tracking. Maintaining channel knowledge at the transmitter requires the use of feedback from the receiver or through the reciprocity principle in a duplex system. If feedback is employed, the receiver must estimate the channel (using training symbols/tones in the transmit signal) and convey the channel state information to the transmitter via a return channel. Alternatively, in a full (frequency or time)-duplex system, the transmitter first learns the "return channel" (i.e., the reverse link) and estimates the channel for the forward link by invoking the reciprocity principle, which guarantees that the transmit and receive channels are identical if the frequency and time (and antennas) of operation on both links are identical, given, of course, that in a duplex system the frequency and time of operation on both links are not identical but close. Maintaining channel knowledge at the transmitter is difficult to implement in practice. For the remainder of this article we focus on the case when the channel is known perfectly to the receiver and is unknown to the transmitter.

4. MIMO CHANNEL CAPACITY

The spectral efficiency of a wireless link is defined as the data rate transmitted per unit bandwidth [bits per second per hertz (bps/Hz)]. The maximum error-free spectral efficiency that can be achieved over a communication link is upper-bounded by the Shannon capacity. In this section we briefly review the Shannon capacity of a MIMO channel for flat fading conditions and then extend the results to frequency selective fading.

The Shannon capacity of a SISO (scalar) channel[1] is given by

$$C = \log_2(1 + \rho\|\mathrm{H}\|^2) \quad \text{bps/Hz} \tag{3}$$

where ρ is the SNR and H is the scalar transfer function. We assume $\mathcal{E}[\mathrm{H}] = 0$ ($\mathcal{E}$ stands for the expectation operator) and $\mathcal{E}[\|\mathrm{H}\|^2] = 1$. As is well known, at high SNR an increase in capacity of 1 bps/Hz is achieved for every 3-dB increase in SNR.

The Shannon capacity of MIMO channels has been derived in [10,11] and is given by[2]

$$C = \log_2\left[\det\left(\mathbf{I}_{M_R} + \frac{\rho}{M_T}\mathbf{H}\mathbf{H}^{\dagger}\right)\right] \quad \text{bps/Hz} \tag{4}$$

where ρ is the SNR defined above for the SISO link and $\mathbf{H}$ is the $M_R \times M_T$ matrix transfer function in (1). We assume $\mathcal{E}[h_{ij}] = 0$ and $\mathcal{E}[\|h_{ij}\|^2] = 1$ (i.i.d. fading model).

Consider a system with an equal number of transmit and receive antennas, $M_T = M_R$. If the channel signatures

[1] A SISO channel can me modeled as a MIMO channel with a 1×1 transfer function.

[2] Here, det ($\mathbf{X}$) stands for the determinant of matrix $\mathbf{X}$. $\mathbf{I}_m$ is the $m \times m$ identity matrix. The superscript † stands for conjugate transpose.

are orthogonal such that $\mathbf{H}\mathbf{H}^{\dagger} = M_T \mathbf{I}_{M_R}$, then the capacity expression in Eq. (4) reduces to

$$C = M_R \log_2(1 + \rho) \text{ bps/Hz} \tag{5}$$

Hence M_R parallel channels are created within the same frequency bandwidth for no additional power expenditure and capacity scales linearly with number of antennas for increasing SNR; that is, the capacity increases by M_R bps/Hz for every 3 dB increase in SNR, leading to a significant capacity advantage. In general, it can be shown that an orthogonal channel of the form described above maximizes the Shannon capacity of a MIMO system. For the i.i.d. fading MIMO channel model discussed in the previous section, the channel realizations become approximately orthogonal when the number of antennas used is very large. When the number of transmit and receive antennas is not equal, $M_T \neq M_R$, the increase in capacity is limited by the minimum of M_T and M_R.

It is important to note that for a time-varying channel, the channel capacity, C, is a random variable whose distribution depends on the channel statistics. To study the capacities of time-varying channels, we can consider the time-averaged channel capacity:

$$\text{Average capacity} = \mathcal{E} \log_2 \left[\det \left(\mathbf{I}_{M_R} + \frac{\rho}{M_T} \mathbf{H}\mathbf{H}^{\dagger} \right) \right] \text{bps/Hz} \tag{6}$$

where $\mathcal{E}$ is the expected value over the distribution of the elements of $\mathbf{H}$. Figure 8 shows the average capacity as a function of the SNR for the i.i.d. fading channel model for different MIMO configurations. It is clear that the average capacity increases with the number of antennas in the system. At very low SNR, the gain in capacity due to multiple antennas is low, but it increases with increasing SNR becoming asymptotically constant.

Since the channel capacity fluctuates with time, it is also useful to define outage capacity for a fading channel.

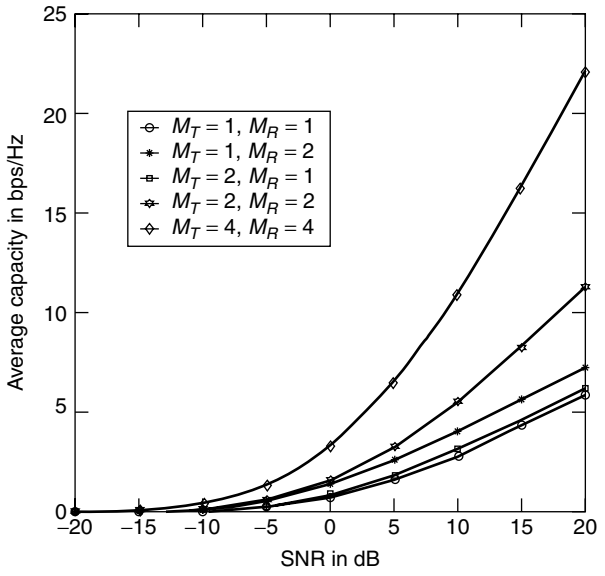


Figure 8. Average capacity for i.i.d. fading MIMO channel model.

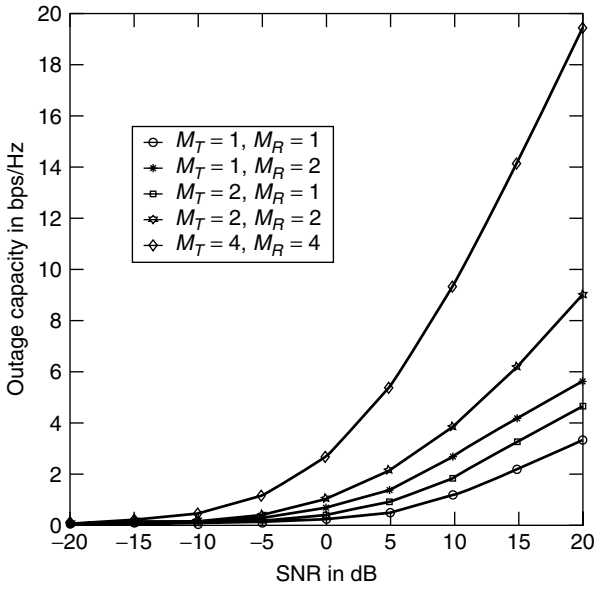


Figure 9. Plot showing 10% outage capacity for i.i.d. fading MIMO channel model.

This is the capacity that is guaranteed at some level of reliability. Thus the 10% outage capacity of a channel is the capacity that is guaranteed 90% of the time. Figure 9 shows the 10% outage capacity as a function of SNR for several MIMO configurations. It is clear that outage capacity also increases with an increasing number of antennas in the system, with better proportionality at larger antenna configurations. This can be attributed to the increased diversity gain at higher values of $M_T \times M_R$.

So far we have restricted our discussion on capacity to a flat fading MIMO channel. The capacity of a frequency-selective fading MIMO channel can be calculated by dividing the frequency band of interest, B, into N narrower flat fading subchannels (Fig. 10). The capacity of the system is then given by the sum of the individual

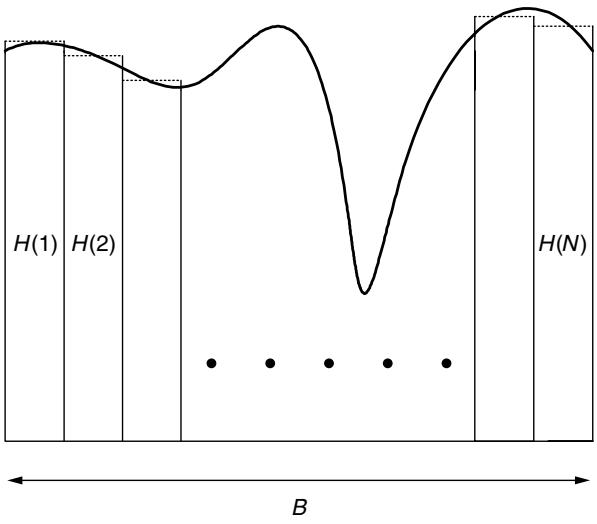


Figure 10. Flat fading approximation of frequency-selective channel.

subchannel capacities

$$C = \frac{1}{N}\sum_{i=1}^{N} \log_2 \det\left(\mathbf{I}_{M_R} + \frac{\rho}{M_T}\mathbf{H}(i)\mathbf{H}(i)^{\dagger}\right) \text{ bps/Hz} \qquad (7)$$

where $\mathbf{H}(i)$ is the channel transfer function corresponding to the ith subchannel. If all subchannels follow the i.i.d. fading MIMO channel model, then averaging over frequency and time reveals that the average capacity of a frequency-selective fading MIMO channel is the same as the average capacity of a frequency-flat fading MIMO channel. However, performance of a frequency-selective fading MIMO channel measured in terms of outage capacity will be better than a frequency-flat fading MIMO channel due to frequency diversity.

The capacity results discussed so far are for the case when the channel is known to the receiver and is unknown to the transmitter. Under this condition, equal power allocation across the transmit antenna array is optimal. If the channel is known to the transmitter, then the optimal transmission strategy from the point of view of maximizing capacity involves allocating possibly unequal amounts of power across the channel modes (corresponding to the singular values of the channel), a technique that is known as *water-filling* [12]. Water-filling may result in an unequal power distribution across the transmit antenna array.

5. MIMO SIGNALING

As described in the previous section, MIMO systems promise much higher spectral efficiency than SISO systems. MIMO systems can also be leveraged to improve the quality of transmission (reduce error rate). Most existing signaling schemes either maximize spectral efficiency (multiplexing mode) or minimize error rate (diversity mode) as will be described next.

5.1. Multiplexing Versus Diversity

We discuss the multiplexing and diversity modes of transmission in the context of a MIMO system with an equal number of transmit and receive antennas. In multiplexing mode, the objective is to maximize the data rate delivered to the receiver. Multiple symbol streams are transmitted to the receiver at the same time as described in Fig. 5. Ideally, the signatures induced at the receive antennas must be orthogonal for spatial multiplexing. The receiver can then perfectly separate the individual symbol streams. On the other hand, if the signatures are not orthogonal, more complex processing at the receiver is required to separate the individual symbol streams. In other words, a low condition number (ideally 1) of $\mathbf{H}$ is preferred for good multiplexing gain. If a channel is not able to support spatial multiplexing because of a high condition number, then data may be delivered to the receiver in diversity mode. Transmit diversity techniques such as space–time coding are employed at the transmitter to extract the spatial diversity in the system. In lieu of the orthogonality requirement of spatial multiplexing, the elements of $\mathbf{H}$ must undergo independent fading for maximum diversity gain. Diversity gain stabilizes the link between transmitter and receiver, improving link reliability.

From the discussion above it is clear that the choice of signaling mode depends on the structure of the MIMO channel. If appropriate channel knowledge is available to the transmitter, it can choose the signaling mode that optimizes performance. This is referred to as *link adaptation*. In general, if the channel is not known to the transmitter, the optimal signaling strategy is a mixture of spatial multiplexing and transmit diversity modes that optimize spectral efficiency and link reliability over the desired range of SNR. Designing efficient, low-complexity receivers for MIMO signaling techniques presents a number of challenges and is a promising area for future research and is described next.

5.2. Receiver Design

In multiplexing mode, each receive antenna observes a superposition of the transmitted signals. The receiver must be able to separate the constituent data streams based on channel knowledge. The separation step determines the computational complexity of the receiver. The problem is similar in nature to the multiuser detection problem in CDMA, and parallels can be drawn between the receiver architectures in these two areas. Maximum-likelihood (ML) detection is optimal but receiver complexity grows exponentially with the number of transmit antennas, making this scheme impractical. Lower-complexity suboptimal receivers include the zero-forcing (ZF) receiver or the minimum mean-square error (MMSE) receiver, the design principles of which are similar to equalization principles for SISO links with intersymbol interference (ISI). An attractive alternative to ZF and MMSE receivers is the V-BLAST algorithm described by Golden et al. [13], which is essentially a successive cancellation technique.

In diversity mode, receiver design is dependent on the diversity signaling technique applied; the most popular is space–time coding. There are two flavors of space–time coding-block codes and trellis codes. Both block codes as well as trellis codes can be designed to extract coding gain and diversity gain. ML receivers and suboptimal receivers similar to those for multiplexing mode have been studied in the context of block codes. The Alamouti scheme [14] is a popular space–time block code for systems with two transmit antennas that uses a simple receiver and extracts maximum diversity gain. Space–time trellis codes are decoded using traditional maximum-likelihood sequence estimation (MLSE) implemented via the Viterbi algorithm. Trellis codes offer better performance than do block codes at the cost of computational complexity. We now briefly describe coding and modulation for MIMO signaling.

5.3. Modulation and Coding for MIMO

MIMO technology is compatible with a wide variety of coding and modulation schemes. In general, the

best performance in achieved by generalizing standard (scalar) modulation and coding techniques to matrix channels. MIMO has been proposed for single-carrier (SC) modulation, direct-sequence code division multiple access (DSCDMA) and orthogonal frequency division multiplexing (OFDM) modulation techniques. MIMO has also been considered in conjunction with single or concatenated coding schemes. Turbo codes and low-density parity codes are currently being studied for MIMO use. As the need for high data rates increases, wireless communication is becoming broadband and there is an increasing trend [15] toward MIMO-OFDM techniques utilizing some version of space–frequency coding with concatenated Reed–Solomon codes.

6. CONCLUDING REMARKS

MIMO wireless communication systems provide significant gains in terms of spectral efficiency and link reliability. These benefits translate to wireless networks in the form of improved coverage and capacity. MIMO communication theory is an emerging area and full of challenging problems. Some promising research areas in the field of MIMO technology include channel estimation, new coding and modulation schemes, low complexity receivers, MIMO channel modeling and network design in the context of MIMO.

Acknowledgments
The authors would like to thank Helmut Bölcskei, Dhananjay Gore, Robert Heath, Sriram Mudulodu, Sumeet Sandhu, and Arak Sutivong for their valuable comments and suggestions. R. Nabar's work was supported by the Dr. T. J. Rodgers Stanford Graduate Fellowship.

BIOGRAPHIES

Arogyaswami J. Paulraj has been a professor at the Department of Electrical Engineering, Stanford University, California, since 1993, where he supervises the Smart Antennas Research Group. This group consists of approximately a dozen researchers working on applications of space-time signal processing for wireless communications networks. His research group has developed many key fundamentals of this new field and has helped shape a worldwide research and development focus on this technology.

Paulraj's research has spanned several disciplines, emphasizing estimation theory, sensor signal processing, parallel computer architectures/algorithms, and space-time wireless communications. His engineering experience includes development of sonar systems, massively parallel computers, and more recently, broadband wireless systems.

He is the author of over 250 research papers and holds 11 patents. Paulraj is a fellow of the Institute of Electrical and Electronics Engineers (IEEE) and a member of the Indian National Academy of Engineering.

Rohit U. Nabar received his B.S. degree in Electrical Engineering in 1998 from Cornell University, Ithaca, New York, and his M.S. degree in electrical engineering in 2000 from Stanford University, California. He is currently a doctoral student in the Smart Antennas Research Group at Stanford University and is the recipient of the Dr. T. J. Rodgers Stanford Graduate Fellowship. His research interests include signal processing and MIMO wireless.

BIBLIOGRAPHY

1. W. C. Y. Lee, *Mobile Communications Engineering*, McGraw-Hill, New York, 1982.
2. W. C. Jakes, *Microwave Mobile Communications*, Wiley, New York, 1974.
3. A. Wittneben, Base station modulation diversity for digital SIMULCAST, *Proc. IEEE VTC*, May 1991, pp. 848–853.
4. N. Seshadri and J. Winters, Two signaling schemes for improving the error performance of frequency-division-duplex (FDD) transmission systems using transmitter antenna diversity, *Int. J. Wireless Inform. Networks* **1**(1): 49–60 (Jan. 1994).
5. J. Guey, M. Fitz, M. Bell, and W. Kuo, Signal design for transmitter diversity wireless communication systems over Rayleigh fading channels, *Proc. IEEE VTC*, 1996, Vol. 1, pp. 136–140.
6. V. Tarokh, N. Seshadri, and A. R. Calderbank, Space-time codes for high data rate wireless communication: Performance criterion and code construction, *IEEE Trans. Inform. Theory* **44**(2): 744–765 (March 1998).
7. V. Tarokh, H. Jafarkhani, and A. R. Calderbank, Space-time block codes from orthogonal designs, *IEEE Trans. Inform. Theory* **45**(5): 1456–1467 (July 1999).
8. U.S. Patent 5,345,599 (1994), A. J. Paulraj and T. Kailath, Increasing capacity in wireless broadcast systems using distributed transmission/directional reception.
9. G. J. Foschini, Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas, *Bell Labs Tech. J.* **1**(2): 41–59 (1996).
10. I. E. Telatar, *Capacity of Multi-antenna Gaussian Channels*, Technical Report BL0112170950615-07TM, AT&T Bell Laboratories, 1995.
11. G. J. Foschini and M. J. Gans, On limits of wireless communications in a fading environment when using multiple antennas, *Wireless Pers. Commun.* **6**(3): 311–335 (March 1998).
12. C. Chuah, D. Tse, and J. M. Kahn, Capacity of multi-antenna array systems in indoor wireless environment, *Proc. IEEE GLOBECOM*, 1998, Vol. 4, pp. 1894–1899.
13. G. D. Golden, G. J. Foschini, R. A. Valenzuela, and P. W. Wolniansky, Detection algorithm and initial laboratory results using the V-BLAST space-time communication architecture, *Electron. Lett.* **35**(1): 14–16 (Jan. 1999).
14. S. M. Alamouti, A simple transmit diversity technique for wireless communications, *IEEE J. Select. Areas Commun.* **16**(8): 1451–1458 (Oct. 1998).
15. H. Bölcskei et al., Fixed broadband wireless: State of the art, challenges and future directions, *IEEE Commun. Mag.* **39**(1): 100–108 (Jan. 2001).

MINIMUM-SHIFT-KEYING

MARVIN K. SIMON
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

1. INTRODUCTION

Minimum-shift-keying (MSK), originally invented by Doelz and Heald as disclosed in a 1961 U.S. Patent [1], is a constant envelope digital modulation that combines both power and bandwidth efficiencies. Signals with constant envelope are desirable when communicating over nonlinear channels, such as those whose transmitter contain a traveling wave tube (TWT) amplifier operated near power saturation, in order to eliminate the occurrence of extraneous spectral sidelobes brought about by amplitude fluctuations. In its native form (also see Hutchinson's 1973 U.S. Patent [2]), MSK is simply a form of binary frequency-shift-keying (BFSK) whose phase is kept continuous from data bit interval to data bit interval and whose modulation index (frequency deviation ratio: the ratio of peak-to-peak frequency deviation to data bit rate) is equal to 0.5. The term *minimum* in this context refers to the fact that, for a given information rate, the 0.5 modulation index corresponds to the minimum frequency shift (and thus the minimum bandwidth) that guarantees orthogonality of the two possible transmitted signals when coherent detection is employed at the receiver which in turn produces maximum power efficiency. Also implicit in the definition of MSK is the fact that the frequency that characterizes the modulation in each bit interval is, as in conventional (not necessarily phase continuous) BFSK, constant over this interval (equivalently, the frequency pulse is a rectangle of duration equal to the bit time).

While at first glance, it might appear that MSK and orthogonal BFSK should have similar performances and behaviors, the fact that the phase is kept continuous in the former introduces memory into the modulation that allows for important differences in the spectral behavior of the transmitted signal as well as the manner in which it can be coherently detected at the receiver. Specifically, the introduction of memory into the modulation produces spectral sidelobes that decay much more rapidly than do those of its conventional binary modulation counterparts. Furthermore, in contrast to a bit-by-bit (bitwise) detector, the deployment of a receiver that exploits the memory introduced at the transmitter offers a power performance more typical of binary *antipodal* signaling than that of binary *orthogonal* signaling.

Although at the time of its introduction MSK had significance in its own right, it gained increased popularity later on when viewed as a special case of a more generic modulation technique referred to as *continuous phase frequency modulation* (CPFM) or more simply *continuous phase modulation* (CPM) whose properties and performance characteristics are well documented in the textbook by Anderson et al. [3]. In particular, CPM allowed for modulation indices other than 0.5, frequency pulse shapes other than rectangular, and frequency pulse durations larger than a single bit time. In fact, it is the distinction between frequency pulses of a single bit duration and those that are longer that accounts for the classification of CPM into *full response* and *partial response* schemes, respectively. Clearly from our discussion above, MSK would fall into the full response CPM category. Furthermore, within the class of full response CPMs, the subclass of schemes having modulation index 0.5 but arbitrary frequency pulse shape resulted in a form of *generalized MSK* [4][1] and included as a special case Amoroso's *sinusoidal FSK* (SFSK) [7] possessing a sinusoidal (raised cosine) frequency pulse shape. Finally, the class of full response schemes with rectangular frequency pulse but arbitrary modulation index is referred to as *continuous phase frequency-shift-keying* (CPFSK) [8] and for all practical purposes served as the precursor to what later became known as CPM itself.

While the primary intent of this article is to focus specifically on the properties and performance of MSK in the form it is most commonly known, the reader should bear in mind that many of these very same characteristics, such as transmitter/receiver implementations, equivalent inphase-quadrature (I-Q) signal representations, and spectral and error probability analysis tools, apply equally well to generalized MSK. Whenever convenient, we shall draw attention to these analogies so as to alert the reader to the generality of our discussions.

In accordance with the above introduction, we begin the mathematical treatment by portraying MSK as a special case of the more general CPM signal whose characterization is given in the next section.

2. THE CONTINUOUS PHASE FREQUENCY MODULATION REPRESENTATION OF MSK

A binary single mode (one modulation index for all transmission intervals) continuous phase modulation (CPM) signal is a constant envelope waveform that has the generic form (see the implementation in Fig. 1).

$$s(t) = \sqrt{\frac{2E_b}{T_b}} \cos(2\pi f_c t + \phi(t, \boldsymbol{\alpha}) + \phi_0), \quad nT_b \le t \le (n+1)T_b \tag{1}$$

where E_b and T_b respectively denote the energy and duration of a bit ($P = E_b/T_b$ is the signal power), and f_c is the carrier frequency. In addition, $\phi(t, \boldsymbol{\alpha})$ is the phase modulation process, which is expressable in the form

$$\phi(t, \boldsymbol{\alpha}) = 2\pi \sum_{i \le n} \alpha_i h q(t - iT_b) \tag{2}$$

where $\boldsymbol{\alpha} = (\ldots, \alpha_{-2}, \alpha_{-1}, \alpha_0, \alpha_1, \alpha_2, \ldots)$ is an independent identically distributed (i.i.d.) binary data sequence with each element taking on equiprobable values ± 1, $h = 2\Delta f T_b$ is the *modulation index* (Δf is the peak frequency deviation

[1] Several other authors [5,6] coined the phrase "generalized MSK" to represent generalizations of MSK other than by pulse shaping.

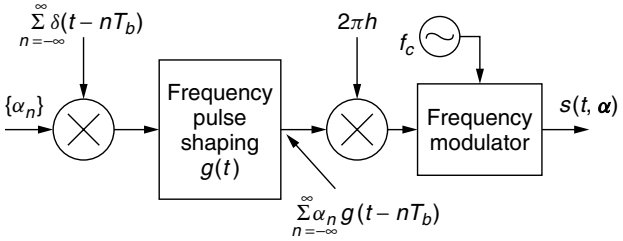


Figure 1. CPM transmitter.

of the carrier), and $q(t)$ is the *normalized phase smoothing response* that defines how the underlying phase $2\pi\alpha_i h$ evolves with time during the associated bit interval. Without loss of generality, the arbitrary phase constant ϕ_0 can be set to zero.

For our discussion here it is convenient to identify the derivative of $q(t)$, namely

$$g(t) = \frac{dq(t)}{dt} \tag{3}$$

which represents the *instantaneous frequency pulse* (relative to the nominal carrier frequency f_c) in the zeroth signaling interval. In view of Eq. (3), the phase smoothing response is given by

$$q(t) = \int_{-\infty}^{t} g(\tau)\, d\tau \tag{4}$$

which, in general, extends over infinite time. For full response CPM schemes, as will be the case of interest here, $q(t)$ satisfies the following:

$$q(t) = \begin{cases} 0, & t \le 0 \\ \frac{1}{2}, & t \ge T_b \end{cases} \tag{5}$$

and thus the frequency pulse $g(t)$ is nonzero only over the bit interval $0 \le t \le T_b$. In view of Eq. (5), we see that the ith data symbol α_i contributes a phase change of $\pi\alpha_i h$ radians to the total phase for all time after T_b seconds of its introduction and thus this fixed phase contribution extends over all future symbol intervals. Because of this overlap of the phase smoothing responses, the total phase in any signaling interval is a function of the present data symbol as well as all of the past symbols and accounts for the *memory* associated with this form of modulation. Thus, in general, optimum detection of CPM schemes must be performed by a *maximum-likelihood sequence estimator* (MLSE) form of receiver [9] as opposed to bit-by-bit detection, which is optimum for memoryless modulations such as conventional BFSK with discontinuous phase.

As previously mentioned, MSK is a full response CPM scheme with a modulation index $h = 0.5$ and a rectangular frequency pulse mathematically described by

$$g(t) = \begin{cases} \dfrac{1}{2T_b}, & 0 \le t \le T_b \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

For SFSK, one of the generalized MSK schemes mentioned in the introduction, $g(t)$ would be a raised cosine pulse given by

$$g(t) = \begin{cases} \dfrac{1}{2T_b}\left[1 - \cos\left(\dfrac{2\pi t}{T_b}\right)\right], & 0 \le t \le T_b \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

The associated phase pulses defined by Eq. (4) are

$$q(t) = \begin{cases} \dfrac{t}{2T_b}, & 0 \le t \le T_b \\ \dfrac{1}{2}, & t \ge T_b \end{cases} \tag{8}$$

for MSK and

$$q(t) = \begin{cases} \dfrac{1}{2T_b}\left[t - \dfrac{\sin 2\pi t/T_b}{2\pi/T_b}\right], & 0 \le t \le T_b \\ \dfrac{1}{2}, & t \ge T_b \end{cases} \tag{9}$$

for SFSK.

Finally, substituting $h = 0.5$ and $g(t)$ of (6) in (1) combined with Eq. (2) gives the CPM representations of MSK and SFSK, respectively, as

$$s_{\text{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}} \cos\left(2\pi f_c t + \frac{\pi}{2T_b}\sum_{i \le n} \alpha_i (t - iT_b)\right), \quad nT_b \le t \le (n+1)T_b \tag{10}$$

and

$$s_{\text{SFSK}}(t) = \sqrt{\frac{2E_b}{T_b}} \cos\left(2\pi f_c t + \frac{\pi}{2T_b} \times \sum_{i \le n} \alpha_i \left[t - iT_b - \frac{\sin 2\pi(t - iT_b)/T_b}{2\pi/T_b}\right]\right), \quad nT_b \le t \le (n+1)T_b \tag{11}$$

both of which are implemented as in Fig. 1 using $g(t)$ of Eqs. (6) or (7) as appropriate.

Associated with MSK (or SFSK) is a *phase trellis* that illustrates the evolution of the phase process with time corresponding to all possible transmitted sequences. For MSK, the phase variation with time is linear [see Eq. (8)] and thus paths in the phase trellis are straight lines with a slope of $\pm\pi/2T_b$. Figure 2 illustrates the MSK phase trellis where the branches are labeled with the data bits that produce the corresponding phase transition. Note that the change in phase over a single bit time is either $\pi/2$ or $-\pi/2$ depending on the polarity of the data bit α_i corresponding to that bit time. Also note that the trellis is *time-varying* in that the phase states (modulo 2π) alternate between 0 and π at even multiples of the bit time and $\pi/2$ and $3\pi/2$ at odd multiples of the bit time. For SFSK the phase trellis would appear as in Fig. 2 with, however, a sinusoidal variation in phase superimposed over the straight line paths. Here again the change in phase over a single bit time would be

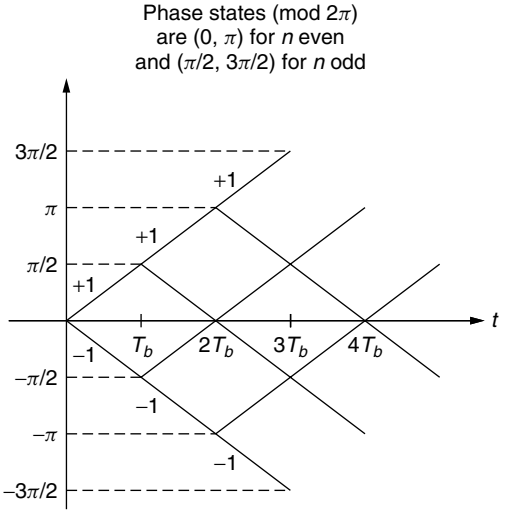


Figure 2. Phase trellis (time-varying) for conventional MSK.

either $\pi/2$ or $-\pi/2$ depending on the polarity of the data bit α_i corresponding to that bit time.

3. EQUIVALENT I-Q REPRESENTATION OF MSK

Although, as stated above, CPM schemes because of their inherent memory require a memory-type of detection, such as MLSE, full response modulations with $h = 0.5$ such as MSK and SFSK can in fact be detected using a memoryless I-Q form of receiver. The reason for this is that for these modulations the transmitter can be implemented in an I-Q form analogous to that of *offset quadrature-phase-shift-keying* (OQPSK). To see this mathematically, we first rewrite the excess phase in the nth transmission interval of the MSK signal in (10) as

$$\phi(t,\boldsymbol{\alpha}) = \frac{\pi}{2T_b}\sum_{i\le n}\alpha_i(t-iT_b) = \alpha_n\frac{\pi}{2T_b}(t-nT_b) + \frac{\pi}{2}\sum_{i\le n-1}\alpha_i = \alpha_n\frac{\pi}{2T_b}t + x_n, \quad nT_b \le t \le (n+1)T_b \tag{12}$$

where $(\pi/2)\sum_{i\le n-1}\alpha_i$ is the accumulated phase at the beginning of the nth transmission interval, which is equal to an odd integer (positive or negative) multiple of $\pi/2$ when n is odd and an even integer (positive or negative) multiple of $\pi/2$ when n is even, and x_n is a phase constant required to keep the phase continuous at the data transition points $t = nT_b$ and $t = (n+1)T_b$. Note also that x_n represents the y-intercept (when reduced modulo 2π) of the path in the phase trellis that represents $\phi(t,\boldsymbol{\alpha})$. In the previous transmission interval, the excess phase is given by

$$\phi(t,\boldsymbol{\alpha}) = \alpha_n\frac{\pi}{2T_b}(t-(n-1)T_b) + \frac{\pi}{2}\sum_{i\le n-2}\alpha_i = \alpha_{n-1}\frac{\pi}{2T_b}t + x_{n-1}, \quad (n-1)T_b \le t \le nT_b \tag{13}$$

For phase continuity at $t = nT_b$, we require that

$$\alpha_n\frac{\pi}{2T_b}(nT_b) + x_n = \alpha_{n-1}\frac{\pi}{2T_b}(nT_b) + x_{n-1} \tag{14}$$

or equivalently

$$x_n = x_{n-1} + \frac{\pi n}{2}(\alpha_{n-1} - \alpha_n) \tag{15}$$

Equation (15) is a recursive relation that allows x_n to be determined in any transmission interval given an initial condition, x_0.

We observe that $(\alpha_{n-1} - \alpha_n)/2$ is a ternary random variable (RV) taking on values $0, +1, -1$ with probabilities $\frac{1}{2}, \frac{1}{4}, \frac{1}{4}$, respectively. Thus, from Eq. (15) when $\alpha_{n-1} = \alpha_n$, $x_n = x_{n-1}$ whereas when $\alpha_{n-1} \ne \alpha_n$, $x_n = x_{n-1} \pm \pi n$. If we arbitrary choose the initial condition $x_0 = 0$, then we see that x_n takes on values of 0 or π (when reduced modulo 2π). Using this fact in (12) and applying simple trigonometry to (10), we obtain

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}[\cos\phi(t,\boldsymbol{\alpha})\cos 2\pi f_c t - \sin\phi(t,\boldsymbol{\alpha})\sin 2\pi f_c t], \quad nT_b \le t \le (n+1)T_b \tag{16}$$

where

$$\begin{aligned}\cos\phi(t,\boldsymbol{\alpha}) &= \cos\left(\alpha_n\frac{\pi}{2T_b}t + x_n\right) \\ &= a_n\cos\frac{\pi}{2T_b}t,\ a_n = \cos x_n = \pm 1 \\ \sin\phi(t,\boldsymbol{\alpha}) &= \sin\left(\alpha_n\frac{\pi}{2T_b}t + x_n\right) \\ &= \alpha_n a_n\sin\frac{\pi}{2T_b}t = b_n\sin\frac{\pi}{2T_b}t, \\ b_n &= \alpha_n\cos x_n = \pm 1\end{aligned} \tag{17}$$

Finally, substituting (17) in (16) gives the I-Q representation of MSK as

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}[a_n C(t)\cos 2\pi f_c t - b_n S(t)\sin 2\pi f_c t], \quad nT_b \le t \le (n+1)T_b \tag{18}$$

where

$$C(t) = \cos\frac{\pi t}{2T_b}, \quad S(t) = \sin\frac{\pi t}{2T_b} \tag{19}$$

are the effective I and Q pulse shapes and $\{a_n\}, \{b_n\}$ as defined in (17) are the effective I and Q binary data sequences.

For SFSK, the representation of Eq. (18) would still be valid with a_n, b_n as defined in (17) but now the effective I and Q pulse shapes become

$$C(t) = \cos\left[\frac{\pi}{2T_b}\left(t - \frac{\sin 2\pi t/T_b}{2\pi/T_b}\right)\right], \quad S(t) = \sin\left[\frac{\pi}{2T_b}\left(t - \frac{\sin 2\pi t/T_b}{2\pi/T_b}\right)\right] \tag{20}$$

To tie the representation of (18) back to that of FSK, we observe that

$$C(t)\cos 2\pi f_c t = \frac{1}{2}\cos\left[2\pi\left(f_c + \frac{1}{4T_b}\right)t\right] + \frac{1}{2}\cos\left[2\pi\left(f_c - \frac{1}{4T_b}\right)t\right]$$
$$S(t)\sin 2\pi f_c t = -\frac{1}{2}\cos\left[2\pi\left(f_c + \frac{1}{4T_b}\right)t\right] + \frac{1}{2}\cos\left[2\pi\left(f_c - \frac{1}{4T_b}\right)t\right] \quad (21)$$

Substituting (21) in (18) gives

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}\left[\left(\frac{a_n + b_n}{2}\right)\cos\left[2\pi\left(f_c + \frac{1}{4T_b}\right)t\right] + \left(\frac{a_n - b_n}{2}\right)\cos\left[2\pi\left(f_c - \frac{1}{4T_b}\right)t\right]\right], \quad nT_b \le t \le (n+1)T_b \quad (22)$$

Thus, when $a_n = b_n (\alpha_n = 1)$, we have

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}\cos\left[2\pi\left(f_c + \frac{1}{4T_b}\right)t\right] \quad (23)$$

whereas when $a_n \neq b_n (\alpha_n = -1)$, we have

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}\cos\left[2\pi\left(f_c - \frac{1}{4T_b}\right)t\right] \quad (24)$$

which establishes the desired connection.

Note from (19) that since $C(t)$ and $S(t)$ are offset from each other by a time shift of T_b seconds, it might appear that $s_{\mathrm{MSK}}(t)$ of (18) is in the form of OQPSK with half sinusoidal pulse shaping.[2] To justify that this is indeed the case, we must examine more carefully the effective I and Q data sequences $\{a_n\}, \{b_n\}$ in so far as their relationship to the input data sequence $\{\alpha_i\}$ and the rate at which they can change. Since the input α_n data bit can change every bit time it might appear that the effective I and Q data bits a_n and b_n can also change every bit time. To the contrary, it can be shown that as a result of the phase continuity constraint of (15), $a_n = \cos x_n$ can change only at the zero crossings of $C(t)$, whereas $b_n = \alpha_n \cos x_n$ can change only at the zero crossings of $S(t)$. Since the zero crossings of $C(t)$ and $S(t)$ are each spaced $2T_b$ seconds apart, then a_n and b_n are constant over $2T_b$-second intervals (see Fig. 3 for an illustrative example). Further noting that the continuous waveforms $C(t)$ and $S(t)$ alternate in sign every $2T_b$ seconds, we can incorporate this sign change in the I and Q data sequences themselves and deal with a fixed *positive* time-limited pulse shape on each

[2] A similar statement can be made for SFSK where the pulse shaping is now described by Eq. (20).

k	α_k	$x_k \pmod{2\pi}$	a_k	b_k	Time interval
0	1	0	1	1	$0 \le t \le T_b$
1	−1	π	−1	1	$T_b \le t \le 2T_b$
2	−1	π	−1	1	$2T_b \le t \le 3T_b$
3	1	0	1	1	$3T_b \le t \le 4T_b$
4	1	0	1	1	$4T_b \le t \le 5T_b$
5	1	0	1	1	$5T_b \le t \le 6T_b$
6	−1	0	1	−1	$6T_b \le t \le 7T_b$
7	1	π	−1	−1	$7T_b \le t \le 8T_b$
8	−1	π	−1	1	$8T_b \le t \le 9T_b$

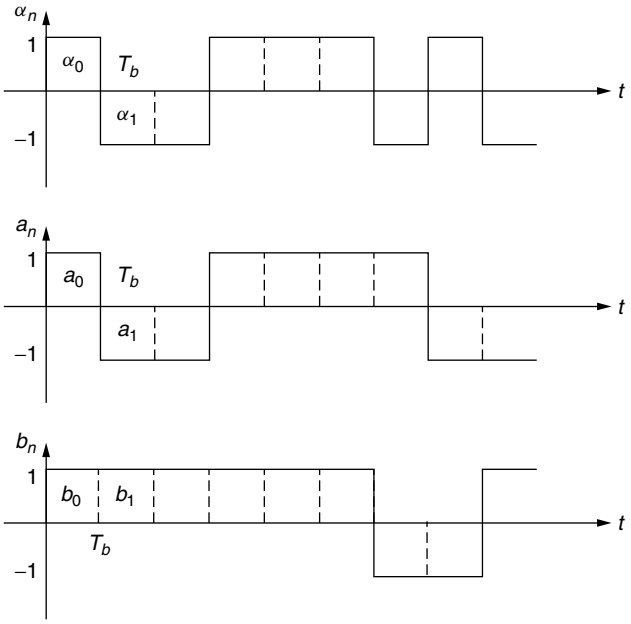


Figure 3. An example of the equivalent I and Q data sequences represented as rectangular pulse streams.

of the I and Q channels. Specifically, defining the pulse shape

$$p(t) = \begin{cases} \sin\dfrac{\pi t}{2T_b}, & 0 \le t \le 2T_b \\ 0, & \text{otherwise} \end{cases} \quad (25)$$

then the I-Q representation of MSK can be rewritten in the form

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}[d_c(t)\cos 2\pi f_c t - d_s(t)\sin 2\pi f_c t] \quad (26)$$

where

$$d_c(t) = \sum_n c_n p(t - (2n-1)T_b), \quad d_s(t) = \sum_n d_n p(t - 2nT_b) \quad (27)$$

with

$$c_n = (-1)^n a_{2n-1}, \quad d_n = (-1)^n b_{2n} \quad (28)$$

To complete the analogy between MSK and sinusoidally pulse shaped OQPSK, we must examine the manner in which the equivalent I and Q data sequences needed in (28) are obtained from the input data sequence $\{\alpha_n\}$. Without going into great mathematical detail, suffice it to say that it can be shown that the sequences

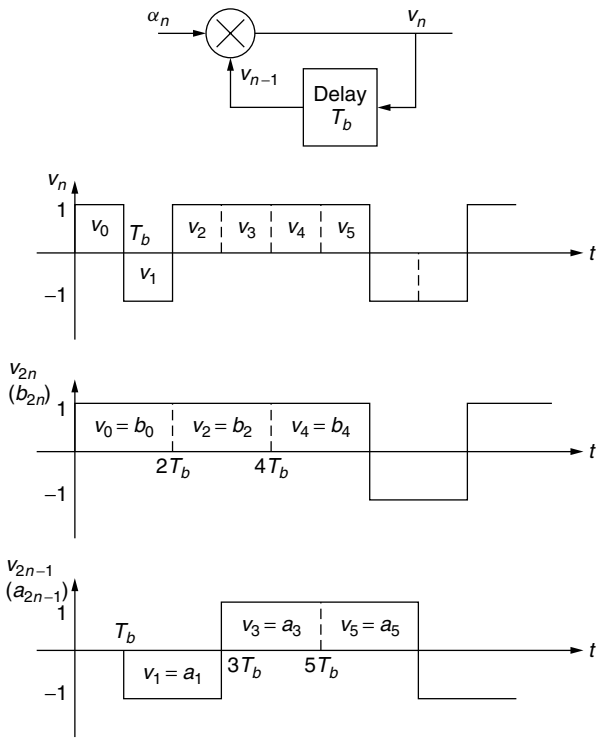


Figure 4. An example of the equivalence between differentially encoded input bits and effective I and Q bits.

$\{a_{2n-1}\}$ and $\{b_{2n}\}$ are the odd/even split of a sequence $\{v_n\}$ which is the *differentially encoded* version of $\{\alpha_n\}$, i.e., $v_n = \alpha_n v_{n-1}$ (see Fig. 4 for an illustrative example). Finally, the I-Q implementation of MSK as described by (27) is illustrated in Fig. 5. As anticipated, we observe that this figure resembles a transmitter for OQPSK except that here the pulse shaping is half-sinusoidal (of symbol duration $T_s = 2T_b$) rather than rectangular, and in addition a differential encoder is applied to the input data sequence prior to splitting it into even and odd sequences each at a rate $1/T_s$. The interpretation of MSK as a special case of OQPSK with sinusoidal pulse shaping along with tradeoffs and comparisons between the two modulations is discussed further in the literature [10,11].

Before concluding this section, we note that the alternative representation of MSK as in (22) can be also expressed in terms of the differentially encoded bits, v_n. In particular

For n odd

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}\left[\left(\frac{v_{n-1}+v_n}{2}\right)\cos\left[2\pi\left(f_c+\frac{1}{4T_b}\right)t\right] - \left(\frac{v_{n-1}-v_n}{2}\right)\cos\left[2\pi\left(f_c-\frac{1}{4T_b}\right)t\right]\right], \quad nT_b \le t \le (n+1)T_b \tag{29a}$$

For n even

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}\left[\left(\frac{v_{n-1}+v_n}{2}\right)\cos\left[2\pi\left(f_c+\frac{1}{4T_b}\right)t\right] + \left(\frac{v_{n-1}-v_n}{2}\right)\cos\left[2\pi\left(f_c-\frac{1}{4T_b}\right)t\right]\right], \quad nT_b \le t \le (n+1)T_b \tag{29b}$$

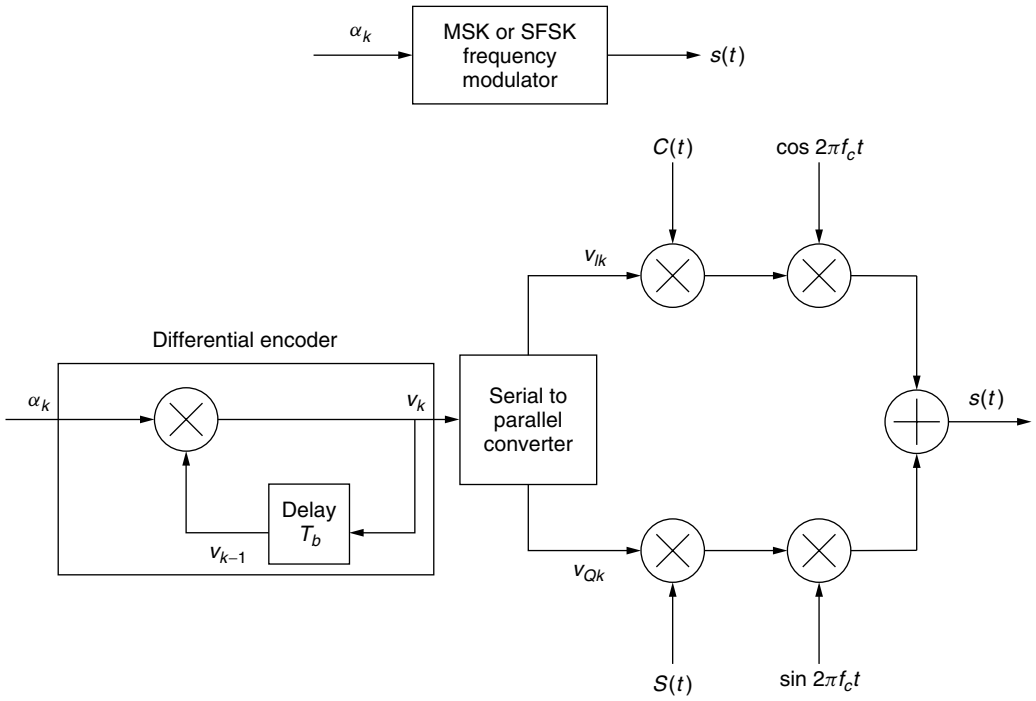


Figure 5. CPM and equivalent I-Q implementations of MSK or SFSK.

Combining these two results, we get

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}\left[\left(\frac{v_{n-1}+v_n}{2}\right)\cos\left[2\pi\left(f_c+\frac{1}{4T_b}\right)t\right] + (-1)^n\left(\frac{v_{n-1}-v_n}{2}\right)\cos\left[2\pi\left(f_c-\frac{1}{4T_b}\right)t\right]\right], \quad nT_b \le t \le (n+1)T_b \tag{30}$$

4. PRECODED MSK

The differential encoder that precedes the I-Q portion of the transmitter in Fig. 5 requires a compensating differential decoder at the receiver following I-Q demodulation and detection (see Fig. 6). Such a combination of differential encoding at the transmitter and differential decoding at the receiver results in a loss in power performance relative to that obtained by conventional OQPSK (this is discussed in more detail later in the article). It is possible to modify MSK to avoid such a loss by first recognizing that the CPM form of modulator in Fig. 1 for implementing MSK can be preceded by the cascade of a differential encoder and a differential decoder without affecting its output (Fig. 7); that is, the cascade of a differential encoder and a differential decoder produces unity transmission, where input = output. Thus, comparing Fig. 7 with Fig. 5, we observe that precoding the CPM form of MSK modulator with a differential *decoder* resulting in what is referred to as *precoded MSK* [9, Chap. 10] will be equivalent to the I-Q implementation of the latter without the differential encoder at its input (see Fig. 8), and thus the receiver for precoded MSK is that of Fig. 6 without the differential decoder at its output. It goes without saying that a similar precoding applied to SFSK would also allow for dispensing with the differential decoder at the output of its I-Q receiver. Finally, we note that both MSK (or SFSK) and its precoded version have identical spectral characteristics and thus for all practical purposes the improvement

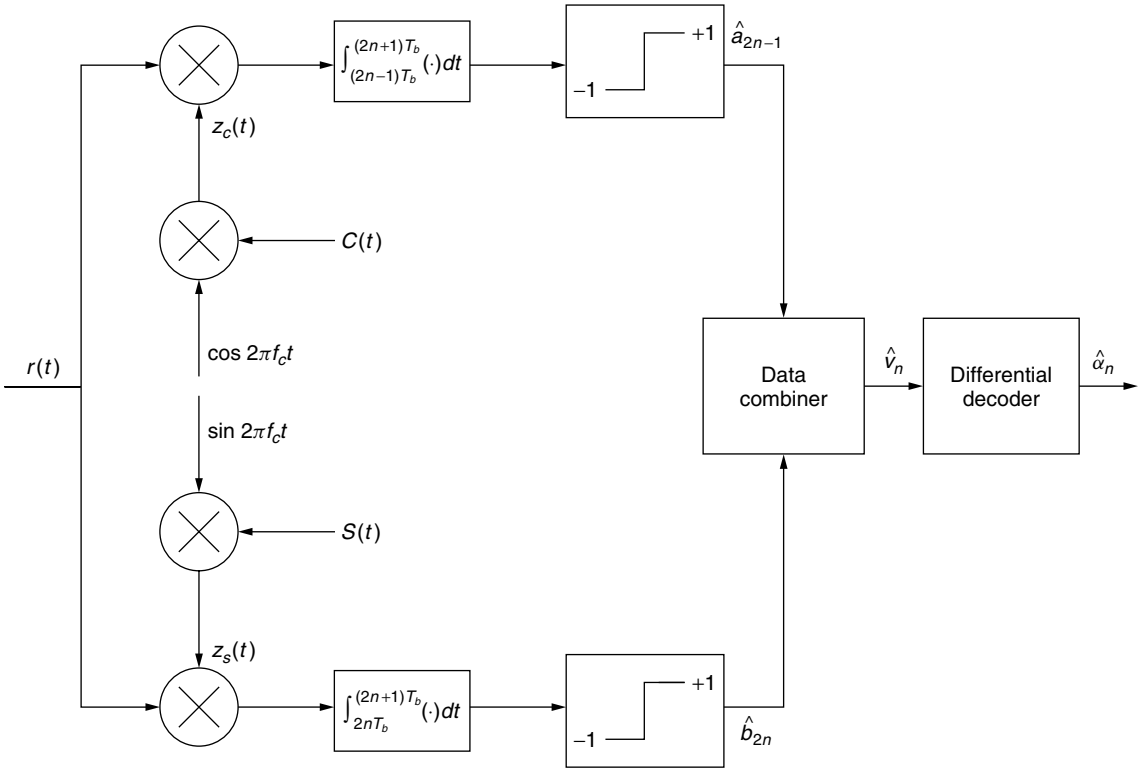


Figure 6. An I-Q receiver implementation of MSK.

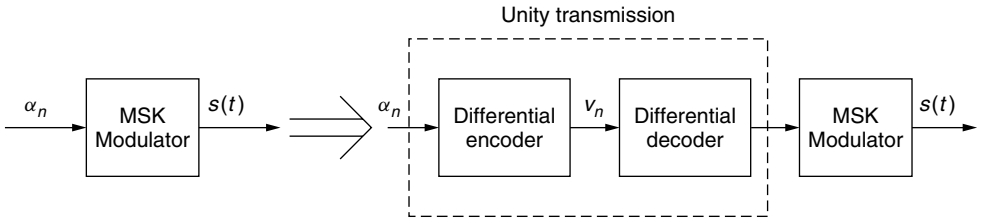


Figure 7. Two equivalent MSK transmitters.

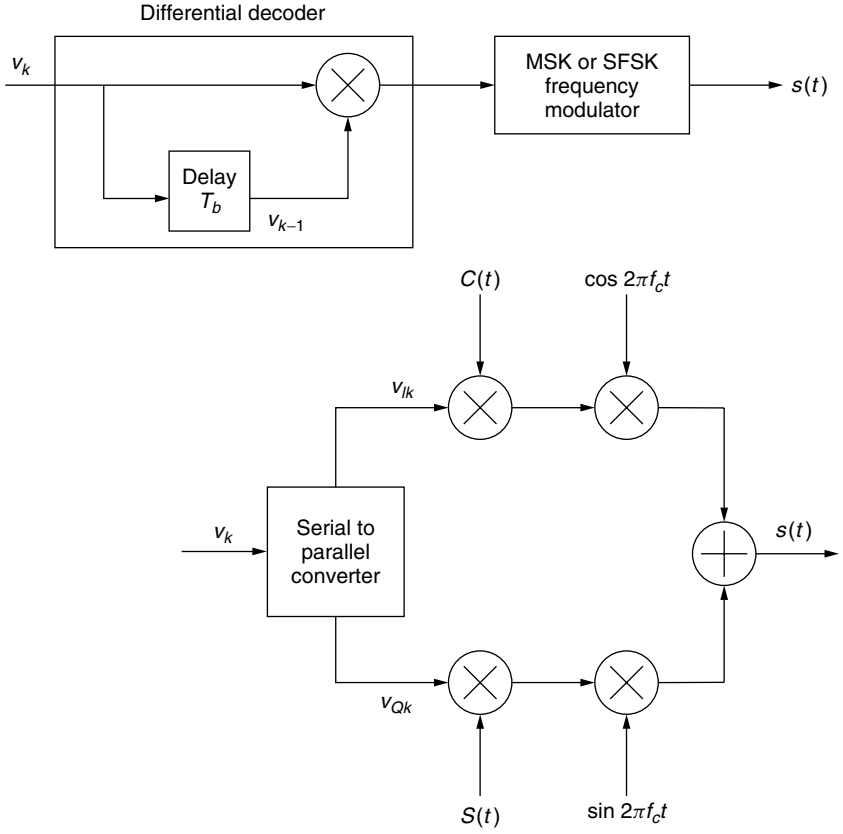


Figure 8. CPM and equivalent I-Q implementations of precoded MSK or SFSK.

in power performance provided by the latter comes at no expense.

5. SPECTRAL CHARACTERISTICS

The ability to express MSK in the offset I-Q form of Eq. (18) allows for simple evaluation of its power spectral density (PSD). In particular, for a generic offset I-Q modulation formed by impressing two lowpass modulations (random pulse trains of rate $1/2T_b$) of equal power and pulse shape on inphase and quadrature carriers:

$$\begin{aligned} s(t) &= Am_I(t)\cos 2\pi f_c t - Am_Q(t)\sin 2\pi f_c t \\ m_I(t) &= \sum_n a_n p(t-2nT_b),\ m_Q(t) = \sum_n b_n p(t-(2n-1)T_b) \end{aligned} \tag{31}$$

the PSD is given by [9, Chap. 2]

$$S_s(f) = \tfrac{1}{4}[G(f-f_c) + G(f+f_c)] \tag{32}$$

where $G(f)$ is the equivalent baseband PSD and is related to the PSD, $S_m(f)$, of $m_I(t)$ or $m_Q(t)$ by

$$G(f) = 2A^2 S_m(f); \quad S_m(f) = \frac{1}{2T_b}|P(f)|^2 \tag{33}$$

with $P(f)$ denoting the Fourier transform of the pulse shape $p(t)$. For MSK, we would have $A = \sqrt{2E_b/T_b}$ and $p(t)$ given by (25) with Fourier transform

$$P(f) = \frac{4T_b}{\pi} e^{-j2\pi f T_b} \frac{\cos 2\pi f T_b}{1 - 16f^2 T_b^2} \tag{34}$$

Substituting (34) in (33) gives the equivalent baseband PSD of MSK as

$$G(f) = \frac{32E_b}{\pi^2} \frac{\cos^2 2\pi f T_b}{(1 - 16f^2 T_b^2)^2} \tag{35}$$

and the corresponding bandpass PSD as [9, Chap. 2]

$$S_s(f) = \frac{8E_b}{\pi^2}\left[\frac{\cos^2 2\pi(f-f_c)T_b}{(1-16(f-f_c)^2T_b^2)^2} + \frac{\cos^2 2\pi(f+f_c)T_b}{(1-16(f+f_c)^2T_b^2)^2}\right] \tag{36}$$

We observe from (35) that the main lobe of the lowpass PSD has its first null at $f = 3/4T_b$. Also, asymptotically for large f, the spectral sidelobes roll off at a rate f^{-4}. By comparison, the equivalent PSD of OQPSK wherein $A = \sqrt{E_b/T_b}$ and $p(t)$ is a unit amplitude rectangular pulse of duration $2T_b$, is given by

$$G(f) = 4E_b \frac{\sin^2 2\pi f T_b}{(2\pi f T_b)^2} \tag{37}$$

whose main lobe has its first null at $f = \frac{1}{2}T_b$ and whose spectral sidelobes asymptotically roll off at a rate f^{-2}. Thus, we observe that while MSK (or precoded MSK) has

a wider main lobe than OQPSK (or QPSK) by a factor of $\frac{3}{2}$, its spectral sidelobes roll off at a rate two orders of magnitude faster. Figure 9 is an illustration of the normalized lowpass PSDs, $G(f)/2E_b$, of MSK and OQPSK obtained from (35) and (37), respectively, as well as that of SFSK, which is given by [9, Chap. 2]

$$G(f) = 2E_b\left[J_0\left(\frac{1}{4}\right)A_0(f) + 2\sum_{n=1}^{\infty} J_{2n}\left(\frac{1}{4}\right)B_{2n}(f) + 2\sum_{n=1}^{\infty} J_{2n-1}\left(\frac{1}{4}\right)B_{2n-1}(f)\right]^2$$

$$A(f) = 2\left(\frac{\sin 2\pi f T_b}{2\pi f T_b}\right), A_0(f) = \frac{1}{2}A\left(f + \frac{1}{4T_b}\right) + \frac{1}{2}A\left(f - \frac{1}{4T_b}\right) = \frac{4}{\pi}\frac{\cos 2\pi f T_b}{1 - 16f^2T_b^2}$$

$$A_{2n}(f) = \frac{1}{2}A\left(f + \frac{2n}{T_b}\right) + \frac{1}{2}A\left(f - \frac{2n}{T_b}\right), \quad (38)$$

$$A_{2n-1}(f) = \frac{1}{2}A\left(f + \frac{2n-1}{T_b}\right) - \frac{1}{2}A\left(f - \frac{2n-1}{T_b}\right)$$

$$B_{2n}(f) = \frac{1}{2}A_{2n}\left(f + \frac{1}{4T_b}\right) + \frac{1}{2}A_{2n}\left(f - \frac{1}{4T_b}\right),$$

$$B_{2n-1}(f) = -\frac{1}{2}A_{2n-1}\left(f + \frac{1}{4T_b}\right) + \frac{1}{2}A_{2n-1}\left(f - \frac{1}{4T_b}\right)$$

$J_n(x) = n$th order Bessel function of the first kind

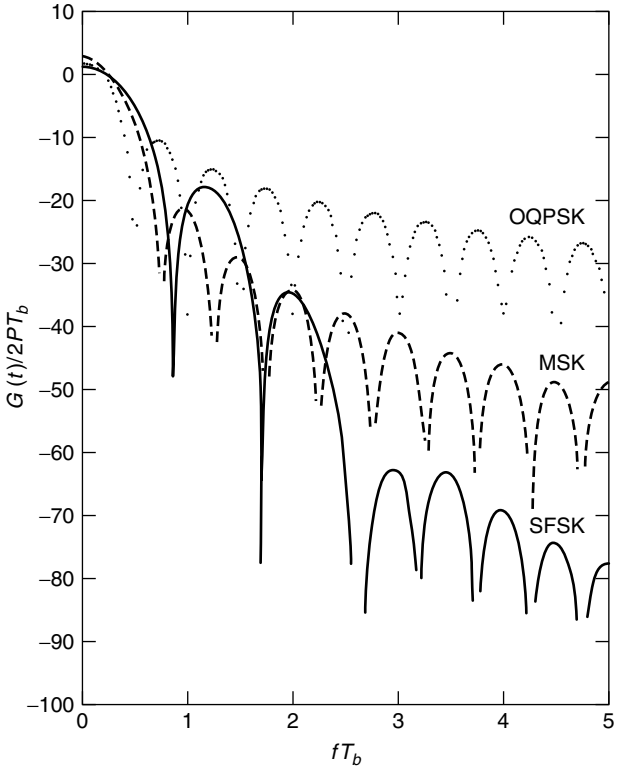


Figure 9. A comparison of the equivalent baseband PSDs of MSK, OQPSK, and SFSK.

whose main lobe is wider than that of MSK but whose spectral sidelobes asymptotically roll off four orders of magnitude faster: at a rate f^{-8}. In fact, for the class of generalized MSK schemes, we can conclude that the smoother we make the shape of the frequency pulse; specifically, the more derivatives that go to zero at the endpoints $t = 0$ and $t = 2T_b$, the wider will be the main lobe but the faster the sidelobes will roll off.

Another way of interpreting the improved bandwidth efficiency that accompanies the equivalent I and Q pulse shaping is in terms of the fractional out-of-band power defined as the fraction of the total power that lies outside a given bandwidth:

$$\eta = 1 - \frac{\int_{-B/2}^{B/2} G(f)\,df}{\int_{-\infty}^{\infty} G(f)\,df} \quad (39)$$

Figure 10 is a plot of the fractional out-of-band power (in decibels) versus BT_b for MSK, OQPSK, and SFSK using the appropriate expression for $G(f)$ as determined from Eqs. (35), (37), and (38), respectively.

6. SERIAL MSK

As an alternative to the parallel I-Q implementations previously discussed, Amoroso and Kivet [12] suggested a serial implementation[3] of MSK which, when the ratio of

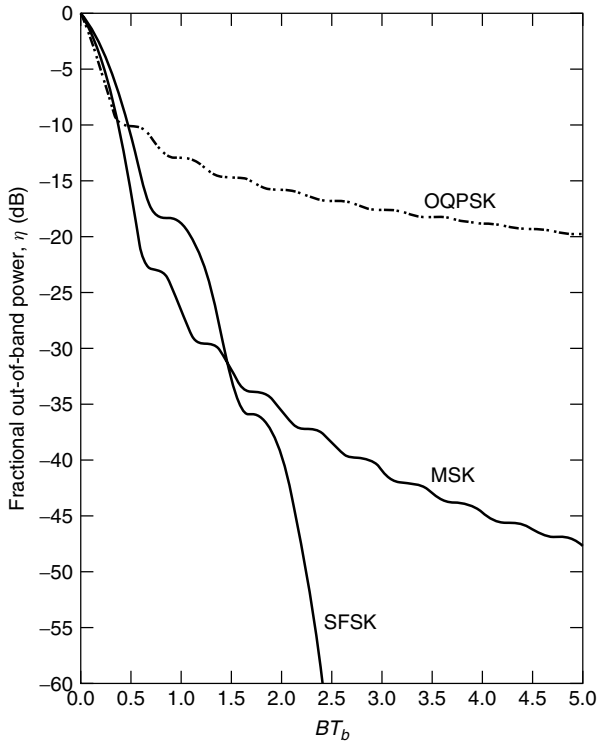


Figure 10. A comparison of the fractional out-of-band power performance of MSK, OQPSK, and SFSK.

[3] Other investigations of serial type implementations appear in [13,14].

carrier frequency to bit rate is high, avoids the requirement for precise relative phasing between the pair of transmitter oscillators needed in the former. In this implementation, MSK is synthesized using a simple biphase modulator that accepts the data in serial form (as opposed to splitting it into I and Q sequences) and as such biphase demodulation and detection are performed at the receiver. The operation of this synthesis method depends on compliance with the original frequency modulation concept of MSK proposed by Doelz and Heald [1] together with additional constraints imposed by Sullivan [15].

Figure 11 is a block diagram of the serial MSK transmitter and receiver. During any T_b-second interval, one of two frequencies f_1 or f_2 is transmitted where for some selected integer n

$$f_1 = \frac{n+1}{2T_b}, \quad f_2 = \frac{n}{2T_b} \tag{40}$$

The carrier frequency f_c is thought of as being midway between f_1 and f_2, namely, $f_c = (f_1 + f_2)/2 = (n + \frac{1}{2})/2T_b$ although it is actually never generated in this implementation. Note that, independent of n, the modulation index $h = (f_1 - f_2)/(1/T_b) = 0.5$ as required for MSK. The operation of the transmitter that synthesizes MSK is as follows. With reference to Fig. 11, a binary rectangular pulse train whose generating sequence is the data sequence $\{\alpha_n\}$ is PSK modulated onto a carrier $c(t) = \cos(2\pi f_2 t + \theta)$ producing the signal

$$x(t) = \sqrt{\frac{2E_b}{T_b}} \sum_{n=-\infty}^{\infty} \alpha_n p(t - nT_b) \cos(2\pi f_2 t + \theta) \tag{41}$$

where $p(t)$ is a unit amplitude pulse of duration T_b seconds and θ is a phase constant to be chosen. This signal is then passed through a lowpass filter with impulse response

$$h_T(t) = \begin{cases} \dfrac{\pi}{T_b} \sin 2\pi f_1 t, & 0 \le t \le T_b \\ 0, & \text{otherwise} \end{cases} \tag{42}$$

Convolving $x(t)$ with $h_T(t)$ gives the filter output in the kth transmission interval as (ignoring high frequency terms at $f_1 + f_2$):

$$\begin{aligned} s(t) = {} & \sqrt{\frac{2E_b}{T_b}} \frac{\pi}{2T_b} \alpha_{k-1} \int_{t-T_b}^{kT_b} \sin[2\pi f_1 t + \theta + 2\pi(f_2 - f_1)\tau]\, d\tau \\ & + \sqrt{\frac{2E_b}{T_b}} \frac{\pi}{2T_b} \alpha_k \int_{kT_b}^{t} \sin[2\pi f_1 t + \theta + 2\pi(f_2 - f_1)\tau]\, d\tau, \\ & kT_b \le t \le (k+1)T_b \end{aligned} \tag{43}$$

Note that $s(t)$ depends only on the 2 data bits α_{k-1} and α_k. Evaluating the integrals and using the fact that $(f_2 - f_1)T_b = 0.5$, we obtain after some simplification

$$\begin{aligned} s(t) = {} & \sqrt{\frac{2E_b}{T_b}} \left\{ -\left(\frac{\alpha_{k-1} + \alpha_k}{2}\right) \cos(2\pi f_2 t + \theta) - (-1)^k \right. \\ & \left. \times \left(\frac{\alpha_{k-1} - \alpha_k}{2}\right) \cos(2\pi f_1 t + \theta) \right\}, \\ & kT_b \le t \le (k+1)T_b \end{aligned} \tag{44}$$

Letting $\theta = \pi$, (44) becomes

$$\begin{aligned} s(t) = {} & \sqrt{\frac{2E_b}{T_b}} \left\{ \left(\frac{\alpha_{k-1} + \alpha_k}{2}\right) \cos 2\pi f_2 t + (-1)^k \left(\frac{\alpha_{k-1} - \alpha_k}{2}\right) \right. \\ & \left. \times \cos 2\pi f_1 t \right\}, kT_b \le t \le (k+1)T_b \end{aligned} \tag{45}$$

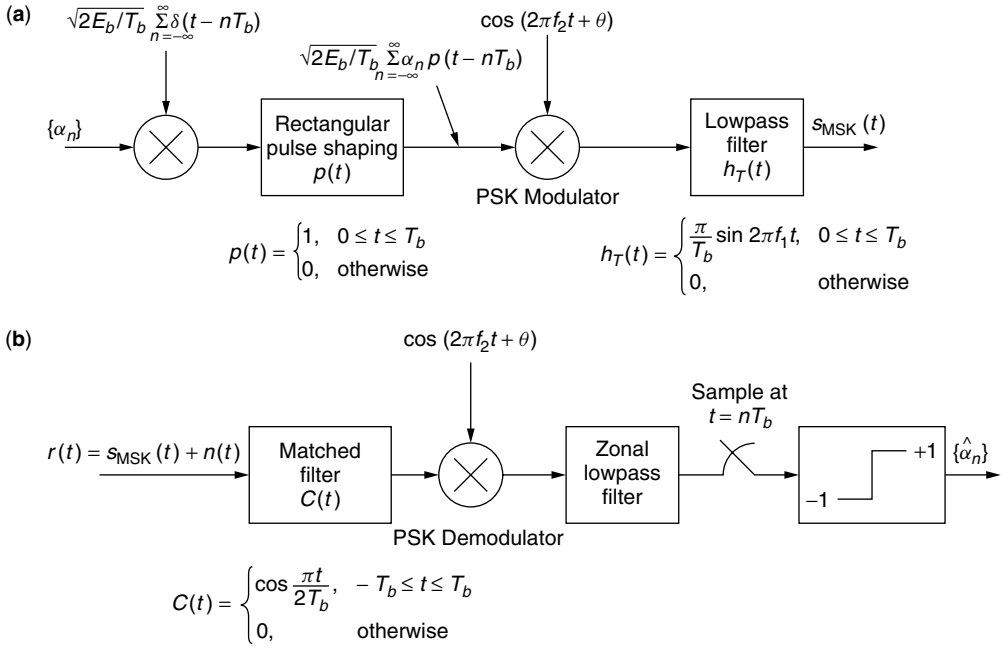


Figure 11. Serial MSK (**a**) transmitter and (**b**) receiver.

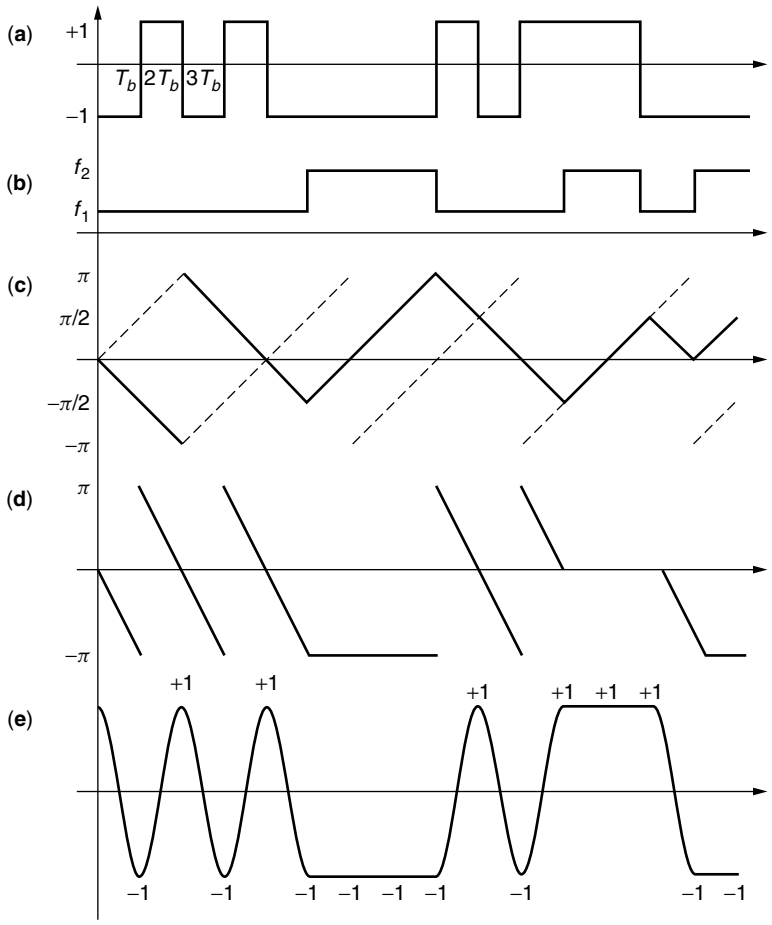


Figure 12. (**a**) Transmitted bit sequence; (**b**) transmitted frequency; (**c**) transmitted excess phase (mod 2π); (**d**) demodulated phase (mod 2π); (**e**) demodulator output (cosine of demodulated phase).

Alternatively, letting $\theta = 0$, we get

$$s(t) = -\sqrt{\frac{2E_b}{T_b}}\left\{\left(\frac{\alpha_{k-1}+\alpha_k}{2}\right)\cos 2\pi f_2 t + (-1)^k\left(\frac{\alpha_{k-1}-\alpha_k}{2}\right)\times\cos 2\pi f_1 t\right\}, kT_b \le t \le (k+1)T_b \quad (46)$$

Comparing Eq. (45) [or (46)] with (30) and noting that $f_1 = f_c - \frac{1}{4T_b}$ and $f_2 = f_c + 1/4T_b$, we see that the serial implementation in Fig. 11 produces a *precoded* MSK signal.[4]

The operation of the serial form of the receiver in Fig. 11 is best described in terms of the series of noise-free waveforms illustrated in Fig. 12 that ignore the presence of the matched filter. Figure 12a is a typical ± 1 data sequence for $\{\alpha_n\}$. Figure 12b shows the transmitted frequency corresponding to this typical data sequence in accordance with (45); that is, frequency f_1 is transmitted when the current bit is different from the previous one and frequency f_2 is transmitted when these 2 bits are the same. The solid line portion of Fig. 12c is the continuous excess (relative to $2\pi f_c t$) phase (reduced modulo 2π) corresponding to the frequency sequence of the second waveform. This is the excess phase of the signal component $s_{\mathrm{MSK}}(t)$ of the received waveform $r(t)$ in Fig. 11b. The excess phase of the PSK demodulation reference is given by $2\pi(f_2 - f_c)t$ and when reduced modulo 2π is illustrated as the dotted line portion of Fig. 12c. The phase of the PSK demodulator output after passing through the zonal lowpass filter (to remove the high frequency carrier term) is given by $\phi(t, \boldsymbol{\alpha}) - 2\pi(f_2 - f_c)t$ and when reduced modulo 2π is illustrated in Fig. 12d. Finally, the actual zonal lowpass filter output is $\cos(\phi(t, \boldsymbol{\alpha}) - 2\pi(f_2 - f_c)t)$, which is illustrated in Fig. 12e. Sampling this waveform at integer multiples of T_b produces a sequence identical to the original data sequence in Fig. 12a. Of course, in the presence of noise, these samples would be noisy, in which case a hard-limiting operation would be used to produce estimates of the data sequence. In the actual implementation of the receiver a matched (to the equivalent I and Q pulse shape) filter would be used which would produce an individual pulse contribution to the recovered bit stream that extends over an interval of $4T_b$ seconds (convolution of a $2T_b$ half-sinusoid with itself). However, it can be shown that the apparent intersymbol interference (ISI) introduced by this broadening of the pulse does not affect the sampled values of the demodulator output (after zonal filtering) and thus no loss

[4] In the Pelchat et al. paper [8], the role of f_1 and f_2 in the transmitter and receiver implementations are reversed with respect to their usage here whose purpose is to maintain consistency with our definition of precoded MSK.

in performance occurs; thus the serial MSK modulation and demodulation system illustrated in Fig. 11 has the same communication efficiency as its parallel counterpart.

Before concluding this section, we note that had we inverted the data sequence in Fig. 12a, the corresponding transmitted frequency sequence of Fig. 12b would remain identical, and likewise the detected data sequence obtained from Fig. 12e would also remain identical. This results in a detected data sequence that is opposite in polarity to the transmitted sequence which implies the presence of a 180° phase ambiguity in the receiver. This type of phase ambiguity is indemic to all binary phase coherent communication systems and as such a means must be provided in the receiver to resolve this ambiguity.

7. CROSS-COUPLED I-Q TRANSMITTER

A variation of the I-Q transmitter discussed in Section 4 is illustrated in Fig. 13 [16–18]. An modulated carrier at frequency f_c is multiplied by a lowpass sinusoidal signal at frequency $1/4T_b$ to produce a pair of unmodulated tones (carriers) at $f_2 = f_c + 1/4T_b$ and $f_1 = f_c - 1/4T_b$. These tones are separately extracted by narrow bandpass filters whose outputs, $s_1(t)$ and $s_2(t)$ are then summed and differenced to produce

$$\begin{aligned} z_c(t) = s_1(t) + s_2(t) &= \frac{1}{2}\cos\left[2\pi\left(f_c - \frac{1}{4T_b}\right)t\right] \\ &\quad + \frac{1}{2}\cos\left[2\pi\left(f_c + \frac{1}{4T_b}\right)t\right] = \cos\left(\frac{\pi t}{2T_b}\right)\cos 2\pi f_c t \\ z_s(t) = s_1(t) - s_2(t) &= \frac{1}{2}\cos\left[2\pi\left(f_c - \frac{1}{4T_b}\right)t\right] \\ &\quad - \frac{1}{2}\cos\left[2\pi\left(f_c + \frac{1}{4T_b}\right)t\right] = \sin\left(\frac{\pi t}{2T_b}\right)\sin 2\pi f_c t \end{aligned} \tag{47}$$

The signals $z_c(t)$ and $z_s(t)$ are respectively multiplied by I and Q data sequences $\{\alpha_I\}$ and $\{\alpha_Q\}$ each at a rate of $\frac{1}{2}T_b$ (and offset from each other by T_b seconds) and then differenced to produce the MSK (actually precoded MSK) output. The advantage of the implementation of Fig. 13 is that the signal coherence and the frequency deviation ratio are largely unaffected by variations in the data rate [17].

8. RIMOLDI'S REPRESENTATION

As stated previously, the conventional CPM implementation of MSK produces a phase trellis that is symmetric about the horizontal axis but time-varying in that the possible phase states (reduced modulo 2π) alternate between $(0, \pi)$ and $(\pi/2, 3\pi/2)$ every T_b seconds. To remove this time-variation of the trellis, Rimoldi [19] demonstrated that CPM with a rational modulation index could be decomposed into the cascade of a memory encoder (finite state machine) and a memoryless demodulator (signal waveform mapper). For the specific case of MSK, Rimoldi's transmitter is illustrated in Fig. 14. Unbalanced (0s and 1s) binary 1 bits, $U_n = (1 - \alpha_n)/2$, are input to a memory one encoder. The current bit and the differentially encoded version of the previous bit (the encoder state) are used to define, via a binary coded decimal (BCD) mapping, a pair of baseband signals (each chosen from a set of four possible waveforms) to be modulated onto I and Q carriers for transmission over the channel. Because of the unbalance of the data, the phase trellis is *tilted* as shown in Fig. 15, but on the other hand it is now *time-invariant*; that is, the phase states (reduced modulo 2π) at all time instants (integer multiples of the bit time) are $(0, \pi)$. This transmitter implementation suggests the use of a simple two-state trellis decoder, which is discussed in the next section dealing with memory receiver structures.

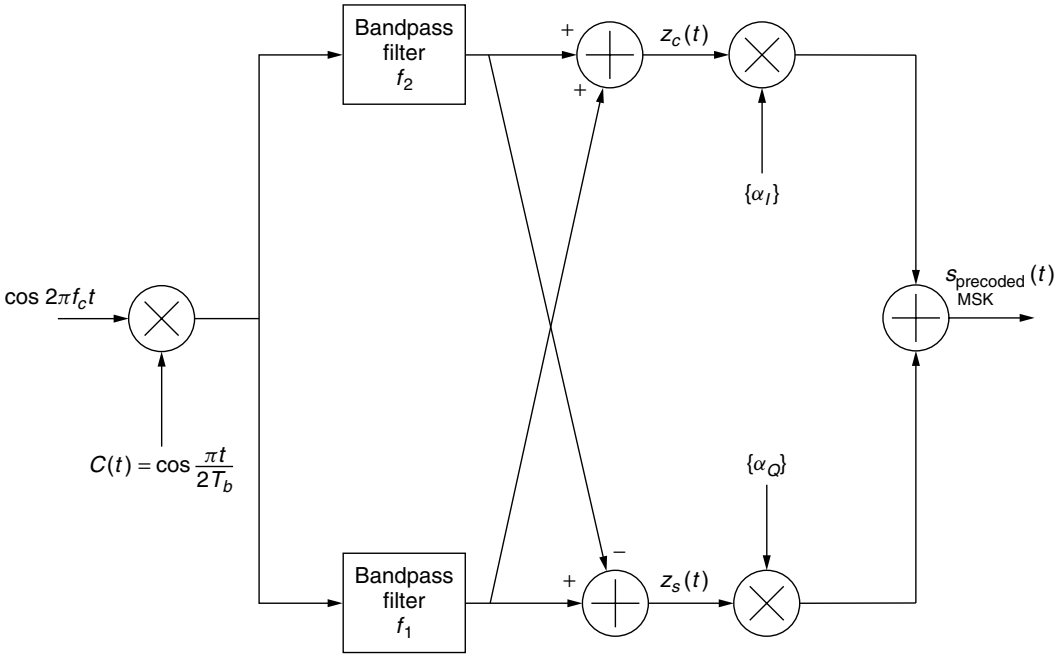


Figure 13. Cross-coupled implementation of precoded MSK.

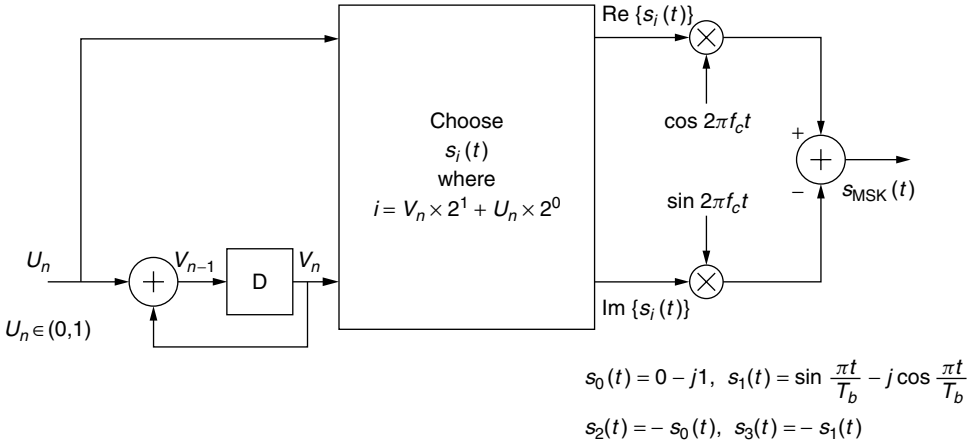


Figure 14. MSK transmitter based on Rimoldi decomposition of CPM.

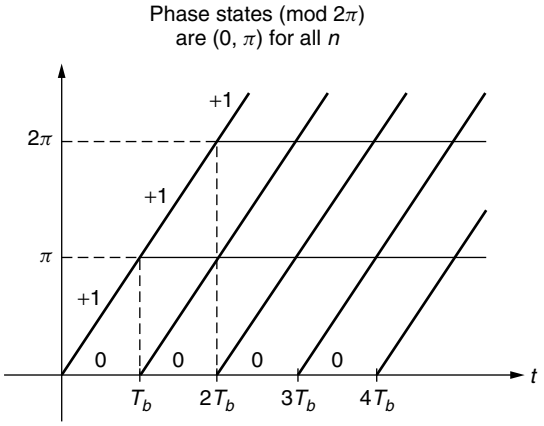


Figure 15. Tilted (time-invariant) phase trellis for Rimoldi's MSK representation.

Rimoldi's representation can also be used to implement precoded MSK. The appropriate transmitter is illustrated in Fig. 16.

9. COHERENT DETECTION

Depending on the particular form used to represent the MSK signal (e.g., CPM, serial or parallel I-Q), many different forms of receivers have been suggested in the literature for performing coherent detection. These various forms fall into two classes: structures based on a memoryless transmitter representation and structures based on a memory transmitter representation. As we shall see, all of these structures, however, are themselves memoryless.

9.1. Structures Based on a Memoryless Transmitter Representation

The two most popular structures for coherent reception of MSK that are based on a memoryless transmitter representation correspond to the parallel I-Q and serial representations and have already been illustrated in Figs. 6 and 11b, respectively. In the case of the former, the received signal plus noise is multiplied by the I and Q "carriers,"[5] $z_c(t)$ and $z_s(t)$, respectively, followed by integrate-and-dump (I&D) circuits of duration $2T_b$ seconds that are timed to match the zero crossings of the I and Q symbol waveforms. The multiplier–integrator combination constitutes a matched filter, which, in the case of additive white Gaussian noise (AWGN) and no ISI,

[5] The word "carrier" here is used to denote the combination (product) of the true carrier and the symbol waveform (clock).

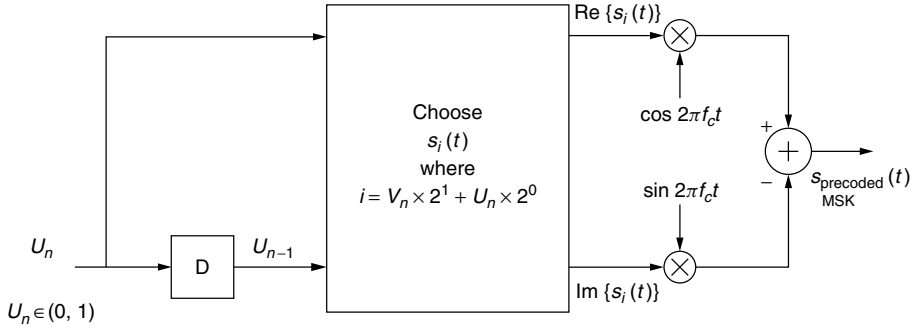


Figure 16. Precoded MSK transmitter based on Rimoldi decomposition of CPM.

results in optimum detection. Means for producing the I and Q demodulation signals $z_c(t)$ and $z_s(t)$ are discussed in the section on synchronization techniques.

9.2. Structures Based on a Memory Transmitter Representation

As noted in Section 7, MSK (or precoded MSK) can be viewed as a cascade of a memory one encoder and a memoryless modulator. As such, a receiver can be implemented on the basis of MLSE detection. For precoded MSK, the appropriate trellis diagram that represents the transitions between states is illustrated in Fig. 17. Each branch of the trellis is labeled with the input bit (0 or 1) that causes a transition and the corresponding waveform (complex) that is transmitted as a result of that transition. The decision metrics based on a two-symbol observation that result in the surviving paths illustrated in Fig. 17 are

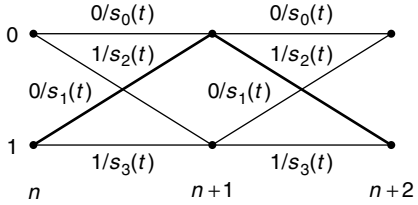
$$\int_{nT_b}^{(n+1)T_b} r(t)s_1(t)\,dt + \int_{(n+1)T_b}^{(n+2)T_b} r(t)s_0(t)\,dt > \int_{nT_b}^{(n+1)T_b} r(t)s_3(t)\,dt + \int_{(n+1)T_b}^{(n+2)T_b} r(t)s_1(t)\,dt \quad (48a)$$

$$\int_{nT_b}^{(n+1)T_b} r(t)s_1(t)\,dt + \int_{(n+1)T_b}^{(n+2)T_b} r(t)s_2(t)\,dt > \int_{nT_b}^{(n+1)T_b} r(t)s_3(t)\,dt + \int_{(n+1)T_b}^{(n+2)T_b} r(t)s_3(t)\,dt \quad (48b)$$

Noting from Fig. 16 that $s_3(t) = -s_0(t)$ and $s_2(t) = -s_1(t)$, (48a) and (48b) can be rewritten as

$$\int_{nT_b}^{(n+1)T_b} r(t)s_0(t)\,dt + \int_{(n+1)T_b}^{(n+2)T_b} r(t)s_0(t)\,dt > -\int_{nT_b}^{(n+1)T_b} r(t)s_1(t)\,dt + \int_{(n+1)T_b}^{(n+2)T_b} r(t)s_1(t)\,dt \quad (49a)$$

$$\int_{nT_b}^{(n+1)T_b} r(t)s_0(t)\,dt + \int_{(n+1)T_b}^{(n+2)T_b} r(t)s_0(t)\,dt > -\int_{nT_b}^{(n+1)T_b} r(t)s_1(t)\,dt + \int_{(n+1)T_b}^{(n+2)T_b} r(t)s_1(t)\,dt \quad (49b)$$



Surviving paths for decoding $\hat{U}_n = 0$ in the interval $(n, n+1)$ assuming state "1" at time n are indicated by heavy lines

Figure 17. Complex baseband trellis.

which are identical and suggest the memoryless receiver illustrated in Fig. 18 [19].[6] Thus we conclude that MSK (or precoded MSK) is a memory one type of trellis-coded modulation (TCM) which can be decoded with a finite (one bit) decoding delay, i.e., the decision on the nth bit can be made at the conclusion of observing the received signal for the $(n+1)$st transmission interval.

Massey [20] suggests an alternative representation of MSK (or precoded MSK) in the form of a single-input two-output sequential transducer followed by an RF selector switch (Fig. 19). For precoded MSK, the sequential transducer implements the ternary sequences $\alpha_k^+ = \frac{1}{2}(\alpha_{k-1} + \alpha_k)$ and $\alpha_k^- = (-1)^k \frac{1}{2}(\alpha_{k-1} - \alpha_k)$ in accordance with Eq. (45). Note as before that α_k^+ is nonzero only when α_k^- is zero and vice versa. The function of the RF selector switch is to select one of the carriers for the signal to be transmitted in each bit interval according to the rule

$$s(t) = \begin{cases} r_2(t) & \text{if } \alpha_k^+ = 1 \\ -r_2(t) & \text{if } \alpha_k^+ = -1 \\ r_1(t) & \text{if } \alpha_k^- = 1 \\ -r_1(t) & \text{if } \alpha_k^- = -1 \end{cases}, \qquad r_i(t) = \sqrt{\frac{2E_b}{T_b}} \cos 2\pi f_i t, \quad i = 1, 2 \quad (50)$$

which represents four mutually exclusive possibilities. This form of modulator has the practical advantage of not requiring addition of RF signals nor RF filtering since there is no actual mixing of the carriers with the modulating signals.

Massey shows that, analogous to Fig. 17, the output of the modulator can be represented by a trellis (Fig. 20) where again each branch is labeled with the input bit

[6] It can be shown that the surviving paths corresponding to being in state "0" at time n leads to the identical decision metric as that in (49a) or (49b).

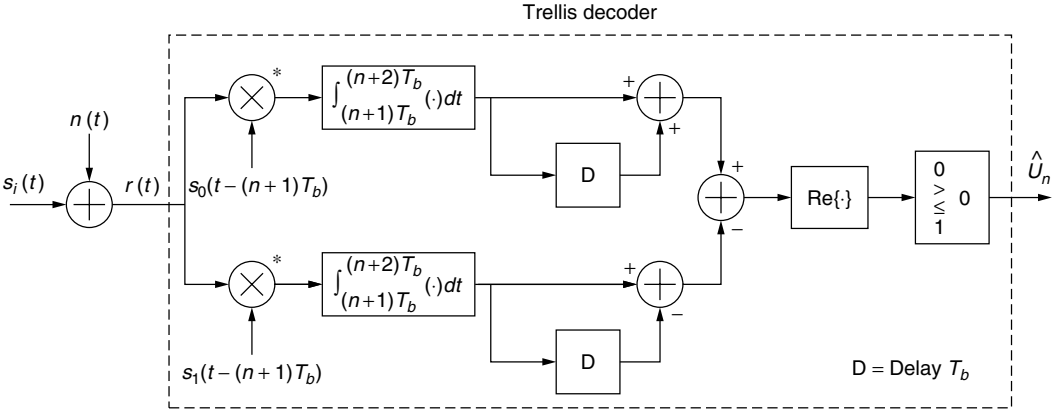


Figure 18. Complex MLSE receiver.

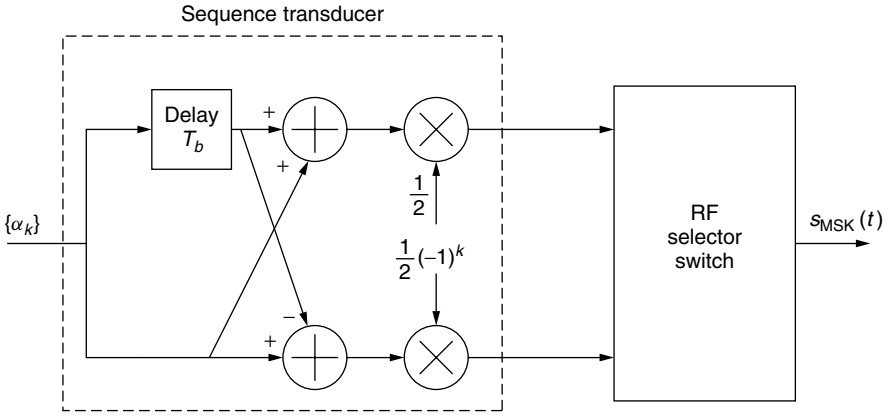


Figure 19. Massey's precoded MSK transmitter.

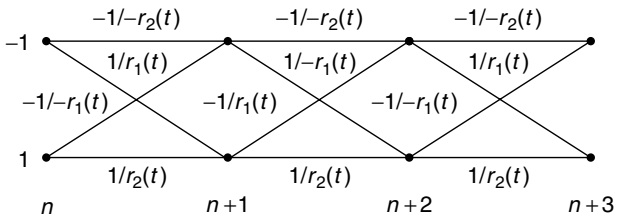


Figure 20. Transmitter output trellis diagram.

and the signal transmitted. Note that the trellis is time-varying (the branch labels alternate with a period of 2). In view of the trellis representation in Fig. 20 the optimum receiver is again an MLSE, which has the same structure as that in Fig. 18, where the complex demodulation signals $s_0(t-(n+1)T_b)$ and $s_1(t-(n+1)T_b)$ are replaced by the real carriers $r_1(t)$ and $r_2(t)$ of (50), the real part of the comparator (difference) output is omitted, and the decision device outputs balanced $+1, -1$ data rather than 0,1 data.

Regardless of the particular receiver implementation employed, the bit error probability (BEP) performance of ideal coherent detection[7] of MSK is given by

$$P_b(E) = \operatorname{erfc}\sqrt{\frac{E_b}{N_0}}\left(1-\frac{1}{2}\operatorname{erfc}\sqrt{\frac{E_b}{N_0}}\right) \tag{51}$$

whereas the equivalent performance of precoded MSK is

$$P_b(E) = \frac{1}{2}\operatorname{erfc}\sqrt{\frac{E_b}{N_0}} \tag{52}$$

which is identical to that of ideal coherent detection of BPSK, QPSK, or OQPSK. Comparing (51) with (52), we observe that the former can be written in terms of the latter as

$$P_b(E)\Big|_{\text{MSK}} = 2P_b(E)\Big|_{\substack{\text{precoded}\\ \text{MSK}}}\left(1-P_b(E)\Big|_{\substack{\text{precoded}\\ \text{MSK}}}\right) \tag{53}$$

which reflects the penalty associated with the differential encoding/decoding operation inherent in MSK but not in precoded MSK as discussed previously. At a BEP of 10^{-5} this amounts to a penalty of approximately a factor of 2 in error probability or equivalently a loss of 0.75 dB in E_b/N_0.

10. DIFFERENTIALLY COHERENT DETECTION

In addition to coherent detection, MSK can be differentially detected [21] as illustrated in Fig. 21. The MSK signal plus noise is multiplied by itself delayed one bit and phase shifted 90°. The resulting product is passed through a lowpass zonal filter that simply removes second harmonics of the carrier frequency terms. Also assumed is that the carrier frequency and data rate are integer related, that is, $f_cT_b = k$ with k integer. Assuming that the MSK signal input to the receiver is in the form of (1) combined with (12):

$$\begin{aligned} s(t) &= \sqrt{\frac{2E_b}{T_b}}\cos\left(2\pi f_c t + \alpha_n\frac{\pi}{2T_b}t + x_n\right)\\ &= \sqrt{\frac{2E_b}{T_b}}\cos\Phi(t,\boldsymbol{\alpha}), \quad nT_b \le t \le (n+1)T_b \end{aligned} \tag{54}$$

then the differential phase $\Delta\Phi \triangleq \Phi(t,\boldsymbol{\alpha}) - \Phi(t-T_b,\boldsymbol{\alpha})$ is given by

$$\Delta\Phi \triangleq -(\alpha_{n-1}-\alpha_n)\frac{\pi}{2}\left(\frac{t}{T_b}-k\right)+\alpha_{n-1}\frac{\pi}{2} \tag{55}$$

where we have made use of the phase continuity relation in (15) in arriving at (55). The mean of the lowpass zonal filter output can be shown to be given by

$$\overline{y(t)} = s(t)s_{90}(t) = \frac{E_b/T_b}{2}\sin\Delta\Phi \tag{56}$$

[7] By "ideal coherent detection" we mean a scenario wherein the local supplied carrier reference is perfectly phase (and frequency) synchronous with the received signal carrier. In Section 10 we explore the practical implications of an imperfect carrier synchronization.

where the "90" subscript denotes a phase shift of 90° in the corresponding signal. Combining (55) and (56), the sampled mean of the lowpass zonal filter output at time $t = (n+1)T_b$ becomes

$$\overline{y((k+1)T_b)} = \frac{E_b/T_b}{2} \sin\left(\alpha_k \frac{\pi}{2}\right) = \alpha_k \frac{E_b/T_b}{2} \qquad (57)$$

which clearly indicates the appropriateness of a hard limiter detector in the presence of noise. Figure 22 illustrates the various waveforms present in the differentially coherent receiver of Fig. 21 for a typical input data sequence.

11. SYNCHRONIZATION TECHNIQUES

In our discussion of coherent reception in Section 6, we implicitly assumed that a means was provided in the receiver for synchronizing the phase of the local demodulation reference(s) with that of the received signal carrier

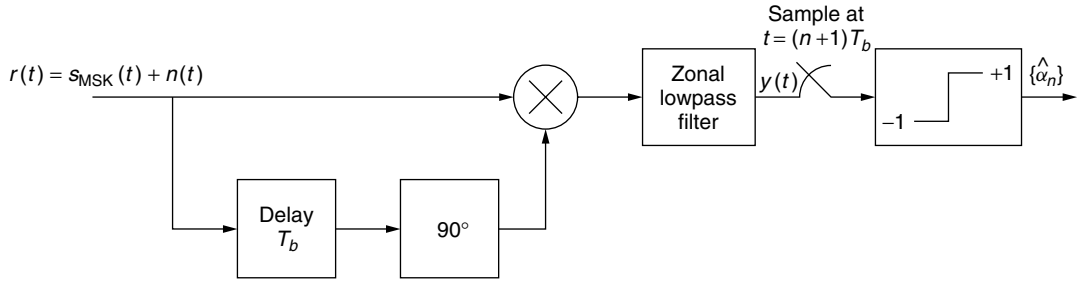


Figure 21. Differentially coherent MSK receiver.

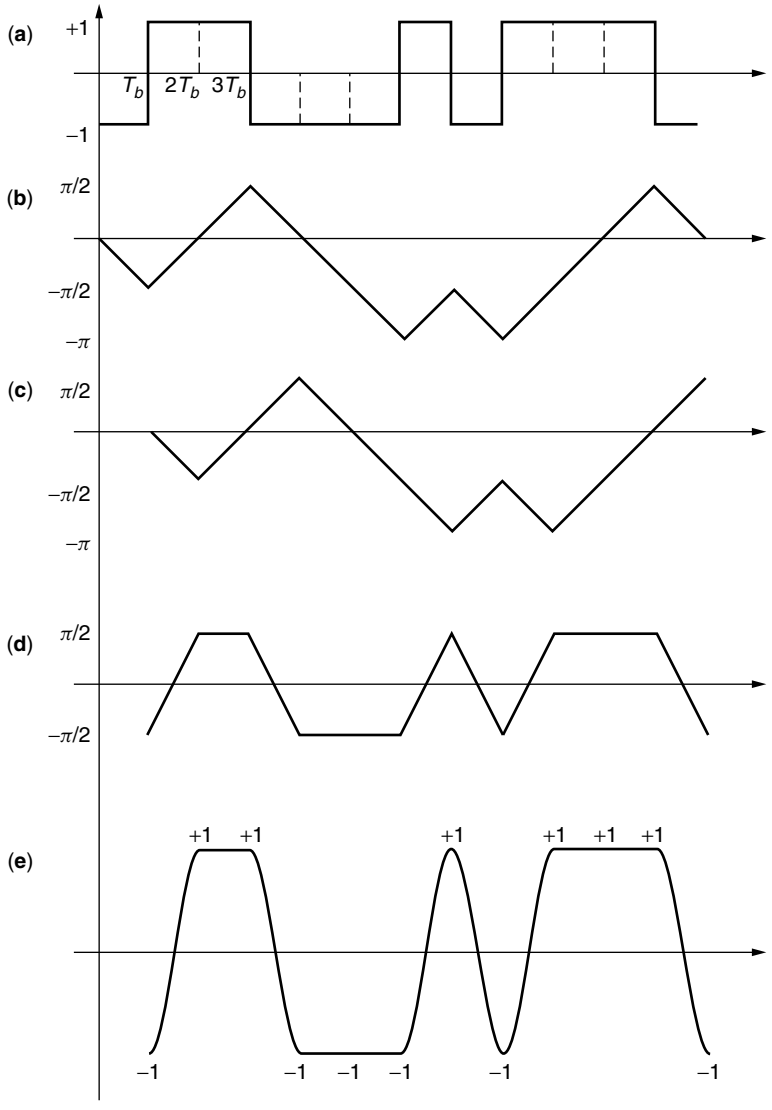


Figure 22. (**a**) Transmitted bit sequence; (**b**) transmitted phase; (**c**) transmitted phase delayed; (**d**) difference phase; (**e**) multiplier output (sine of difference phase).

and also for time synchronizing the I&D circuits. Here we discuss several options for implementing such means.

One form of combined carrier and clock recovery which is synergistic with the transmitter form in Fig. 13 was originally proposed by DeBuda [22,23].[8] With reference to Fig. 23, the received MSK signal is first squared to produce an FSK signal at twice the carrier frequency and with twice the modulation index, i.e., $h = 1$, which is known as *Sunde's FSK* [24]. Whereas the MSK signal has no discrete (line) spectral components, after being squared it has strong spectral components at $2f_1$ and $2f_2$ which can be used for synchronization. In fact, Sunde's FSK has 50% of its total power in these two line components (the other 50% of the total power is in a discrete line component at DC). To demonstrate this transformation from continuous to discrete spectrum, we square the MSK signal form in (30), which gives

$$\begin{aligned} s^2_{\mathrm{MSK}}(t) &= \frac{2E_b}{T_b}[(v_n^+)^2\cos^2 2\pi f_2 t + (v_n^-)^2\cos^2 2\pi f_1 t \\ &\quad + 2v_n^+ v_n^- \cos 2\pi f_2 t \cos 2\pi f_1] \\ &= \frac{2E_b}{T_b}\left[\frac{1}{2} + \frac{1}{2}(v_n^+)^2\cos 4\pi f_2 t + \frac{1}{2}(v_n^-)^2\cos 4\pi f_1 t\right], \\ &\qquad v_n^+ = \frac{v_{n-1}+v_n}{2},\ v_n^- = (-1)^n\left(\frac{v_{n-1}-v_n}{2}\right) \end{aligned} \tag{58}$$

where we have made use of the fact that since either v_n^+ or v_n^- is always equal to zero, then $v_n^+ v_n^- = 0$. Also, either $(v_n^+)^2 = 1$ and $(v_n^-)^2 = 0$ or vice versa, which establishes (58) as a signal with only discrete line components. The components at $2f_1$ and $2f_2$ are extracted by bandpass filters (in practice, phase-locked loops) and then frequency divided to produce $s_1(t) = \frac{1}{2}\cos 2\pi f_1 t$ and $s_2(t) = \frac{1}{2}\cos 2\pi f_2 t$. The sum and difference of these two signals produce the reference "carriers" $z_c(t) = C(t)\cos 2\pi f_c t$ and $z_s(t) = S(t)\sin 2\pi f_c t$, respectively, needed in Fig. 6. Finally, multiplying $s_1(t)$ and $s_2(t)$ and lowpass filtering the result produces $\frac{1}{8}\cos 2\pi t/2T_b$ (a signal at half the bit rate), which provides the desired timing information for the I&Ds in Fig. 6.

Another joint carrier and timing synchronization scheme for MSK was derived by Booth [25] in the form of a closed loop motivated by the maximum a posteriori (MAP) estimation of carrier phase and symbol timing (Fig. 24). The resulting structure (Fig. 24a) is an overlay of two MAP estimation I-Q closed loops — one typical of a carrier synchronization loop assuming known symbol timing (Fig. 24b) and one typical of a symbol timing loop assuming known carrier phase (Fig. 24c). In fact, the carrier synchronization component loop is identical to what would be obtained for sinusoidally pulse-shaped OQPSK.

Finally, many other synchronization structures have been developed for MSK and conventional (single modulation index) binary CPM, which, by definition, would also be suited to MSK. A sampling of these is given in the literature [26–32]. In the interest of brevity, however, we do not discuss these here. Instead the interested reader is referred to the cited references for the details.

12. FURTHER EXTENSIONS

As alluded to earlier, MSK is just one special case of a class of full response CPM modulations with modulation index 0.5, in particular, it is one that possesses a rectangular frequency pulse shape. We have also at times referred to another modulation in this class, namely,

[8] DeBuda also referred to MSK, in conjunction with his self-synchronizing circuit, as "fast FSK (FFSK)" which at the time was the more popular terminology in Canada.

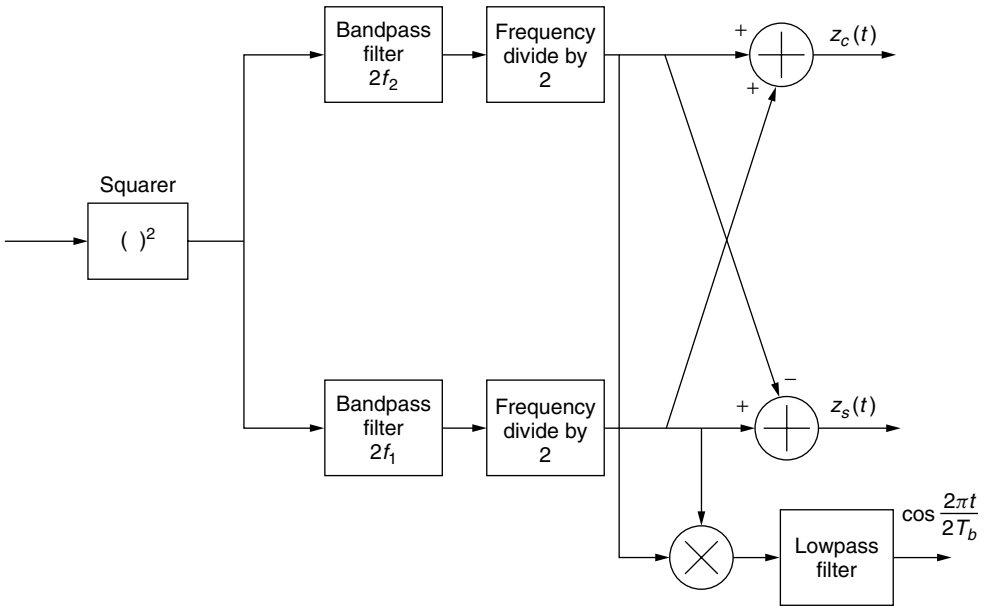


Figure 23. DeBuda's carrier and symbol synchronization scheme.

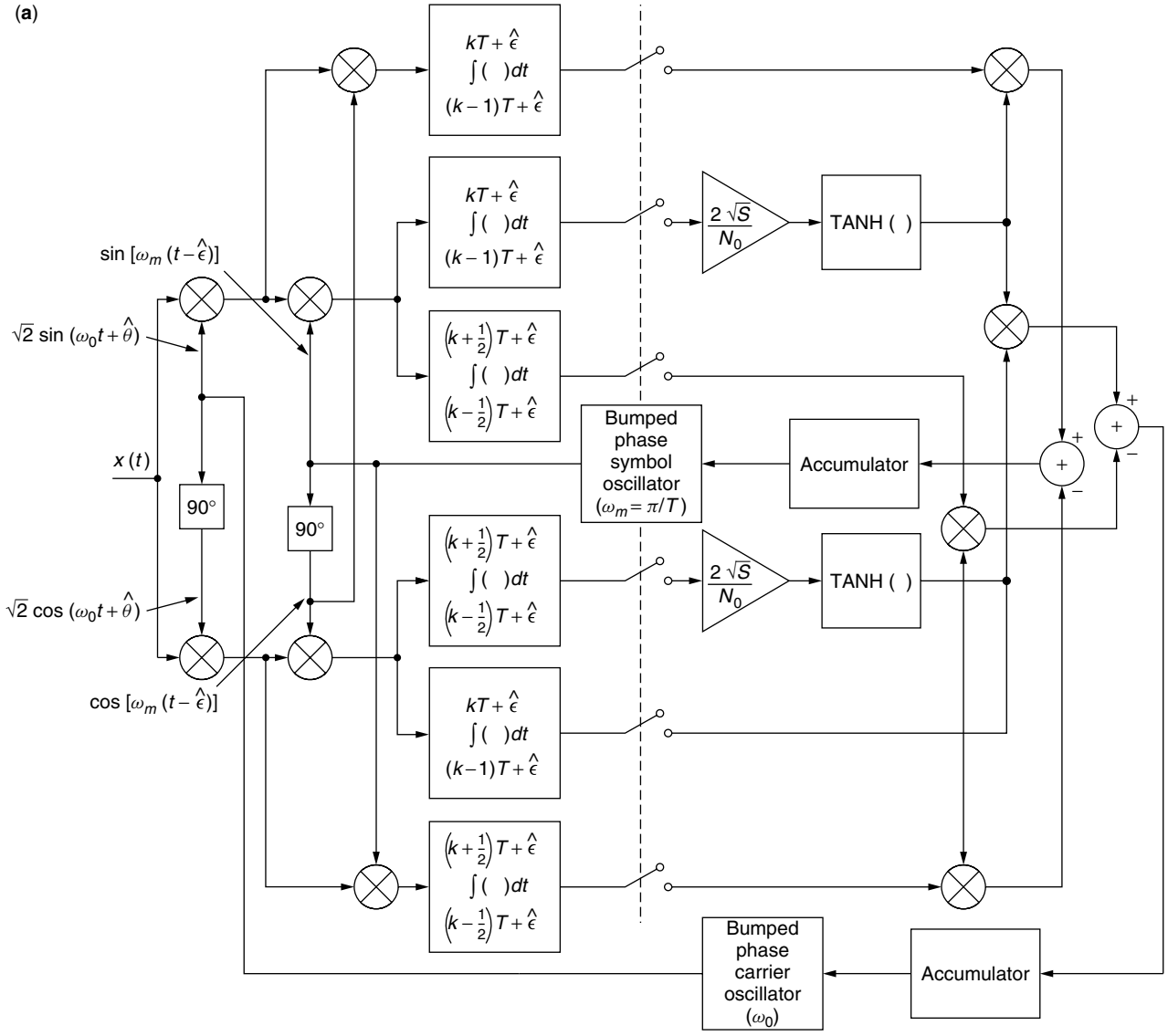


Figure 24. (**a**) Joint carrier and symbol MAP estimation loop for MSK modulation; (**b**) same (carrier synchronization component); (**c**) same (symbol synchronization component).

SFSK, which possesses a raised cosine pulse shape. Since the particular frequency pulse shape selected does not effect the power efficiency (error probability performance) of the system, provided an appropriate matched filter is used in the receiver, then the choice of a pulse shape is made primarily based on spectral efficiency considerations. In this regard, many other modulations in this class have been suggested in the literature. Here we briefly summarize these and provide the appropriate references for readers wishing to explore these in more detail.

Reiffen and White [33] proposed a generalization of MSK called *continuous shift keying* (CSK), in which the instantaneous frequency and perhaps higher derivatives of the phase, as well, are continuous. For CSK the effective inphase channel amplitude pulse shape (of duration $2T_b$) takes the form

$$C(t)=\frac{1}{\sqrt{T_b}}\cos\phi(t), \quad -T_b\le t\le T_b \tag{59}$$

where

$$\phi(t)=\begin{cases}\pm\dfrac{\pi}{2}, & t=-T_b\\ \text{arbitrary}, & -T_b<t<0\\ 0, & t=0\\ \pm\dfrac{\pi}{2}\pm\phi_0(t-T_b), & 0<t<T_b\\ \pm\dfrac{\pi}{2}, & t=T_b\end{cases} \tag{60}$$

It can be shown that the spectral efficiency of this class of schemes is directly related to the smoothness of the derivatives at the endpoints of the $2T_b$-second interval.

Rabzel and Pasupathy [34] proposed a pulse shape characterized by Eq. (59) but with

$$\phi(t)=\frac{\pi t}{2T_b}-\frac{1}{n}\sum_{i=1}^{M}K_i\left[\sin\frac{2\pi nt}{T_b}\right]^{2i-1},$$

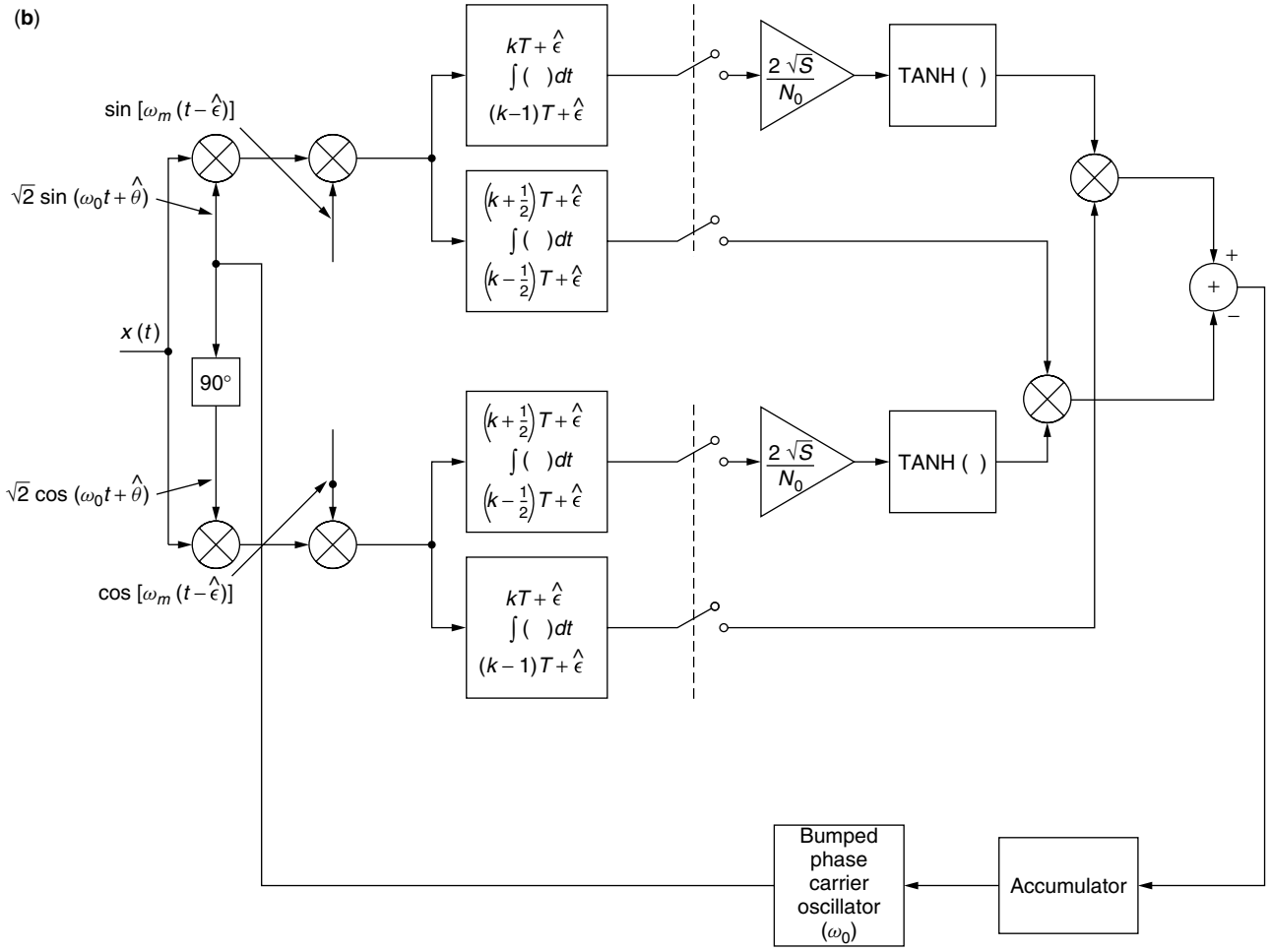


Figure 24. (*Continued*)

$$K_i = \frac{(2i-2)!}{2^{2i-2}[(i-1)!]^2(2i-1)} \tag{61}$$

where $n = 1, 2, 3, \ldots,$ and M is an integer parameter to be selected. The PSD of this class of generalized MSK modulations asymptotically rolls off as $f^{-(4M+4)}$. Special cases are $M = 0$ (MSK) and $M = 1, n = 1$ (SFSK).

Bazin [35] proposed a pulse shape also characterized by (59) but with

$$\phi(t) = \frac{\pi t}{2T_b} - \sum_{k=1}^{N'} A_k \sin \frac{2\pi k t}{T_b}, \quad N' \geq \frac{N}{2} \tag{62}$$

where N is an integer such that all the phase pulse function has all its derivatives up to the Nth order equal to zero at its endpoints (as such the PSD asymptotically rolls of as $f^{-(2N+4)}$) and the A_k coefficients are the solution of the linear system

$$\left.\frac{d^i s(t)}{dt^i}\right|_{t=\pm T_b} = 0, \quad i = 1, 2, \ldots, N \tag{63}$$

It can be shown that Rabzel's pulse format is a subclass of Bazin's. Aside from the well-known special cases of MSK and SFSK, a variation of the latter called DSFSK is the special case corresponding to $N = 4$, $N' = 2$ with coefficients $A_1 = \frac{1}{3}$, $A_2 = -\frac{1}{24}$ which, from Eq. (62), yields the phase pulse

$$\phi(t) = \frac{\pi t}{2T_b} - \frac{1}{3}\sin\frac{2\pi t}{T_b} + \frac{1}{24}\sin\frac{4\pi t}{T_b} \tag{64}$$

In accordance with the above, the PSD for this modulation scheme asymptotically rolls off as f^{-12}. Finally, the connection between CSK and MSK, SFSK, and the generalizations suggested in [34] and [35] was pointed out by Cruz [36].

As an alternative to the choice of a frequency pulse for shaping the transmitted PSD, one can accomplish spectral efficiency by introducing correlation into the transmitted data sequence via a precoder. In Section 3 we considered a simple differential decoder as a precoder in Fig. 8; however, such a precoder has no effect whatsoever on the PSD. The authors of Refs. 37–41 applied correlative encoding (duobinary modulation) [42] to MSK with the intention of obtaining spectral improvement with minimum sacrifice in power performance. The duobinary encoder is illustrated in Fig. 25 and at first glance resembles a differential decoder (see Fig. 8). However, one

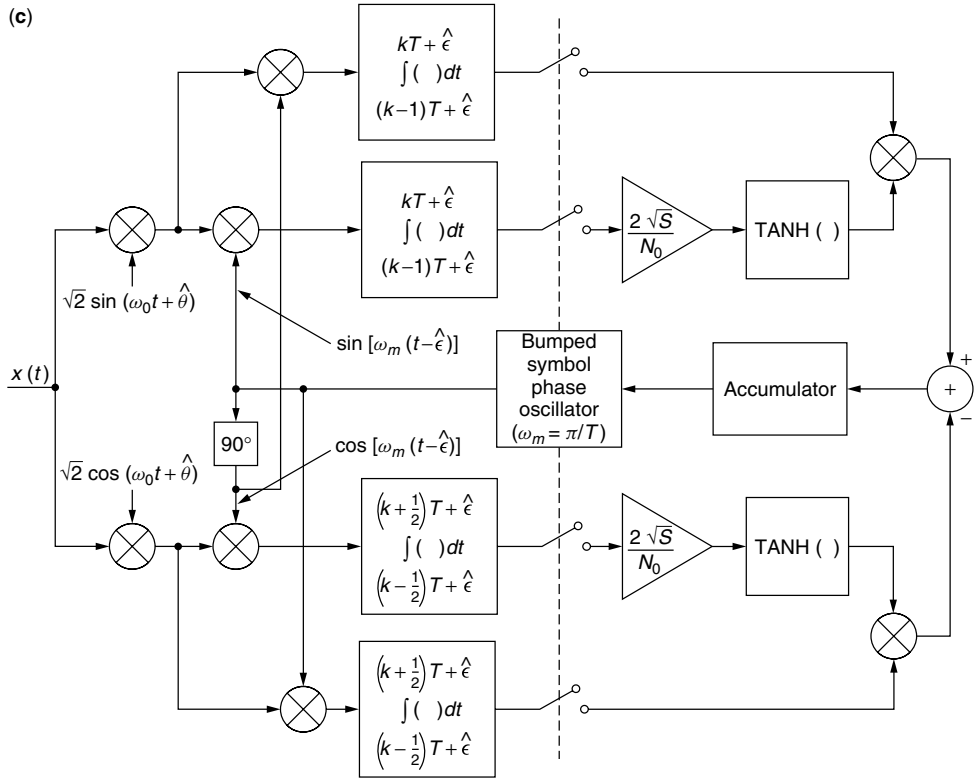


Figure 24. (*Continued*)

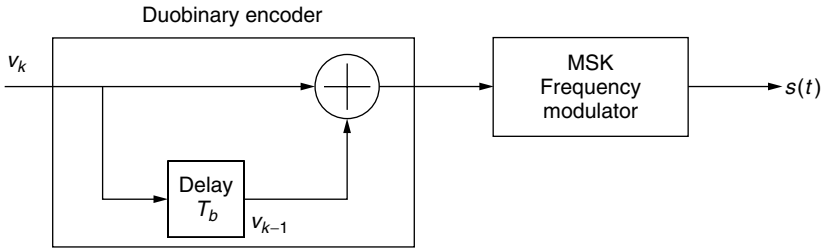


Figure 25. Duobinary encoded MSK.

should note that the multiplier in the latter is replaced by a summer in the former, which results in a ternary $(+1, 0, -1)$ rather than a binary $(+1, -1)$ output.

The last extension of MSK worth mentioning, which again attempts to trade improved bandwidth efficiency for a decrease in power efficiency, is the extension to multiple level pulses, known as *multiple amplitude MSK* (MAMSK) [43]. The simplest way of envisioning this modulation is to consider the I-Q representation of MSK where the input data $\{\alpha_n\}$ now takes on levels $\pm 1, \pm 3, \ldots, \pm M-1$. In this regard, MAMSK can be viewed as a form of offset QAM with sinusoidal pulse shaping [44]. Of course, since the modulation now contains multiple amplitudes, the modulation is no longer constant envelope but rather occupies a discrete number of envelope levels.

Before concluding this chapter we wish to point out to the reader a well-written and simple-to-read tutorial article on MSK by Pasupathy [45] that covers the basics of what is discussed here and provides a valuable set of additional references on the subject as of the time of that publication that include some of the early applications in commercial communication systems.

BIOGRAPHY

Dr. Marvin K. Simon is currently a principal scientist at the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, where for the last 34 years he has performed research as applied to the design of NASA's deep-space and near-earth missions resulting in the issuance of nine patents and 23 NASA Tech Briefs. His research interests are modulation and demodulation, synchronization techniques for space, satellite and radio communications, trellis-coded modulation, spread spectrum and multiple access communications, and communication over fading channels. In the past, Dr. Simon also held a joint appointment with the Electrical Engineering Department at Caltech.

He has published over 160 papers and 10 textbooks on the above subjects. His work has also appeared as chapters in several other textbooks. He is the corecipient of

the 1988 Prize Paper Award in Communications of the IEEE Transactions on Vehicular Technology for his work on trellis-coded differential detection systems and also the 1999 Prize Paper of the IEEE Vehicular Technology Conference for his work on switched diversity. He is a fellow of the IEEE and a fellow of the IAE. Among his awards are the NASA Exceptional Service Medal, NASA Exceptional Engineering Achievement Medal, IEEE Edwin H. Armstrong Achievement Award, and most recently the IEEE Millennium Medal all in recognition of outstanding contributions to the field of digital communications and leadership in advancing this discipline.

BIBLIOGRAPHY

1. U.S. Patent 2,977,417 (March 28, 1961), M. L. Doelz and E. T. Heald, Minimum-shift data communication system.
2. U.S. Patent 3,731,233 (May 1, 1973), W. M. Hutchinson, Minimum shift keying modulating apparatus.
3. J. B. Anderson, T. Aulin, and C.-E. Sundberg, *Digital Phase Modulation*, Plenum Press, New York, 1986.
4. M. K. Simon, A generalization of MSK-Type signaling based upon input data symbol pulse shaping, *IEEE Trans. Commun.* **COM-24**(8): 845–856 (Aug. 1976).
5. P. Galko and S. Pasupathy, Generalized MSK, *Proc. IEEE Int. Electrical, Electronics, Conf. Exposition*, Toronto, Ontario, Canada, Oct. 5–7, 1981.
6. I. Korn, Generalized MSK, *IEEE Trans. Inform. Theory* **IT-26**(2): 234–238 (March 1980).
7. F. Amoroso, Pulse and spectrum manipulation in the minimum (frequency) shift keying (MSK) format, *IEEE Trans. Commun.* **COM-24**(3): 381–384 (March 1976).
8. M. G. Pelchat, R. C. Davis, and M. B. Luntz, Coherent demodulation of continuous phase binary FSK signals, *Proc. Int. Telemetry Conf.* Washington, DC, 1971.
9. M. K. Simon, S. M. Hinedi, and W. C. Lindsey, *Digital Communication Techniques: Signal Design and Detection*, Prentice-Hall, Upper Saddle River, NJ, 1995.
10. F. Amoroso and J. A. Kivett, Simplified MSK signaling technique, *IEEE Trans. Commun.* **25**(4): 433–441 (April 1977).
11. H. R. Mathwich, J. F. Balcewicz, and M. Hecht, The effect of tandem band and amplitude limiting on the E_b/N_0 performance of minimum (frequency) shift keying (MSK), *IEEE Trans. Commun.* **COM-22**(10): 1525–1540 (Oct. 1974).
12. S. A. Gronemeyer and A. L. McBride, MSK and offset QPSK modulation, *IEEE Trans. Commun.* **COM-24**(8): 809–820 (Aug. 1976).
13. R. E. Ziemer, C. R. Ryan, and J. R. Stilwell, Conversion and matched filter approximations for serial minimum-shift keyed modulation, *IEEE Trans. Commun.* **COM-30**(3): 495–509 (March 1982).
14. S. M. Ryu and C. K Un, A simple method for MSK modulation and demodulation, *Proc. IEEE* **73**(11): 1690–1691 (Nov. 1985).
15. W. A. Sullivan, High-capacity microwave system for digital data transmission, *IEEE Trans. Commun.* **COM-20**(P. I): 466–470 (June 1972).
16. D. M. Brady, A constant envelope digital modulation technique for millimeter-wave satellite system, *ICC'74 Conf. Record*, Minneapolis, MN, June 1974, p. 36C-1.
17. D. P. Taylor, A high speed digital modem for experimental work on the communications technology satellite, *Can. Elect. Eng. J.* **2**(1): 21–30 (1977).
18. R. M. Fielding, H. L. Berger, and D. L. Lochhead, Performance characterization of a high data rate MSK and QPSK channel, *ICC'77 Conf. Record*, Chicago, IL, June 1977, pp. 3.2.42–3.2.46.
19. B. E. Rimoldi, A decomposition approach to CPM, *IEEE Trans. Inform. Theory* **IT-34**: 260–270 (May 1988).
20. J. L. Massey, A generalized formulation of minimum shift keying modulation, *ICC'80 Conf. Record*, Seattle, WA, June 1980, pp. 26.5.1–26.5.5.
21. T. Masamura, S. Samejima, Y. Morihiro, and H. Fuketa, Differential detection of MSK with nonredundant error correction, *IEEE Trans. Commun.* **COM-27**(6): 912–918 (June 1979).
22. R. DeBuda, The Fast FSK modulation system, *ICC'71 Conf. Record*, Montreal, Canada, June 1971, pp. 41-25–45-27.
23. R. DeBuda, Coherent demodulation of frequency-shift-keying with low deviation ratio, *IEEE Trans. Commun.* **COM-20**(3): 429–435 (June 1972).
24. W. R. Bennett and S. O. Rice, Spectral density and autocorrelation functions associated with binary frequency shift keying, *Bell Syst. Tech. J.* **42**: 2355–2385 (Sept. 1963).
25. R. W. Booth, An illustration of the MAP estimation method for deriving closed-loop phase tracking topologies: the MSK signal structure, *IEEE Trans. Commun.* **COM-28**(8): 1137–1142 (Aug. 1980).
26. S. J. Simmons and P. J. McLane, Low-complexity carrier tracking decoders for continuous phase modulations, *IEEE Trans. Commun.* **COM-33**(12): 1285–1290 (Dec. 1985).
27. J. Huber and W. Liu, Data-aided synchronization of coherent CPM receivers, *IEEE Trans. Commun.* **40**(1): 178–189 (Jan. 1992).
28. M. Moeneclaey and I. Bruyland, The joint carrier and symbol synchonizability of continuous phase modulated waveforms, *ICC'86 Conf. Record*, Vol. 2, Toronto, Canada, June 1986, pp. 31.5.1–31.5.5.
29. A. N. D'Andrea, U. Mengali, and R. Reggiannini, A digital approach to clock recovery in generalized minimum shift keying, *IEEE Trans. Vehic. Technol.* **39**: 227–234 (Aug. 1990).
30. A. N. D'Andrea, U. Mengali, and M. Morelli, Multiple phase synchronization in continuous phase modulation, in *Digital Signal Processing 3*, Academic Press, New York, 1993, pp. 188–198.
31. U. Lambrette and H. Meyr, Two timing recovery algorithms for MSK, *ICC'94 Conf. Record*, New Orleans, LA, May 1994, pp. 918–992.
32. A. N. D'Andrea, U. Mengali, and M. Morelli, Symbol timing estimation with CPM modulation, *IEEE Trans. Commun.* **44**(10): 1362–1371 (Oct. 1996).
33. B. Reiffen and B. E. White, On low crosstalk data communication and its realization by continuous shift keyed modulation schemes, *IEEE Trans. Commun.* **COM-26**(1): 131–135 (Jan. 1978).

34. M. Rabzel and S. Pasupathy, Special shaping in minimum shift keying (MS)-type signals, *IEEE Trans. Commun.* **COM-26**(1): 189–195 (Jan. 1978).

35. B. Bazin, A class of MSK baseband pulse formats with sharp spectral roll-off, *IEEE Trans. Commun.* **COM-27**(5): 826–829 (May 1979).

36. J. R. Cruz, A note on spectral shaping of minimum-shift-keying-type signals, *Proc. IEEE* **68**(8): 1035–1036 (Aug. 1980).

37. G. J. Garrison, A power spectral density analysis for digital FM, *IEEE Trans. Commun.* **COM-23**(11): 1228–1243 (Nov. 1975).

38. F. De Jager and C. B. Dekker, Tamed frequency modulation: A novel method to achieve spectrum economy in digital transmission, *IEEE Trans. Commun.* **COM-26**(50): 534–542 (May 1978).

39. S. Gupta and S. Elnoubi, Error rate performance of coded MSK with discriminator detection in land mobile communication system, *Proc. Int. Communications and Computer Exposition*, Los Angeles, CA, Nov. 1980, pp. 120–124.

40. S. Gupta and S. Elnoubi, Error rate performance of duobinary coded MSK and TFM with differential detection in land mobile communication systems, *31st Vehicular Technology Conf. Record*, Washington, DC, April 1981.

41. S. Gupta and S. Elnoubi, Error rate performance of noncoherent detection of duobinary coded MSK and TFM in mobile radio communication systems (with S. Elnoubi), *IEEE Trans. Vehic. Technol.* **VT-30**: 62–76 (May 1981).

42. S. Pasupathy, Correlative coding: a bandwidth efficient signaling scheme, *IEEE Commun. Mag.* **17**(4): 4–11 (July 1977).

43. W. J. Weber, P. H. Stanton, and J. T. Sumida, A bandwidth compressive modulation system using multiamplitude minimum shift-keying (MAMSK), *IEEE Trans. Commun.* **COM-26**(5): 543–551 (May 1978).

44. M. K. Simon, An MSK approach to offset QASK, *IEEE Trans. Commun.* **COM-24**(8): 921–923 (Aug. 1976).

45. S. Pasupathy, Minimum shift keying: a spectrally efficient modulation, *IEEE Commun. Mag.* **17**(4): 14–22 (July 1979).

MOBILE RADIO COMMUNICATIONS

RODGER E. ZIEMER
University of Colorado
Colorado Springs, Colorado

WILLIAM H. TRANTER
R. MICHAEL BUEHRER
Virginia Tech
Blacksburg, Virginia

THEODORE S. RAPPAPORT
The University of Texas at Austin
Austin, Texas

1. THE EARLY HISTORY OF WIRELESS COMMUNICATIONS

Guglielmo Marconi's development and commercialization of wireless telegraphy in the 1890s marked the beginning of wireless communications. While nineteenth century researchers such as Volta, Hertz, and Tesla experimented with the electrostatic and inductive components of electromagnetic fields, Marconi accidentally discovered that a radiation field could be launched from an appropriately designed antenna, thereby allowing reliable propagation over great distances. In April 1901, Marconi successfully demonstrated wireless transmission across the Atlantic Ocean. The work of Marconi spawned a century of research and commercial activity that produced the AM radio, FM radio, television, land mobile radio, cellular radio, wireless data networks, and satellite communications.

Even during the early days of wireless communications, mobile communications was of interest. In 1902, Ernest Rutherford and his assistant, Howard Barnes, developed a wireless telegraphy system to communicate with moving trains. One of the pioneering practical uses of mobile wireless telegraphy was ship-to-ship and ship-to-shore communications in the early decades of the twentieth century. Wireless communications played an important role in rescue operations when the Titanic sank in 1912. Although many recognized the commercial potential of wireless communications, the dawn of the twentieth century witnessed a number of skeptics. For example, J. J. Thompson, who was to receive the Nobel Prize in physics in 1906, remarked that wireless communications was not likely to ever be of real commercial use [1]. In addition, Ernest Rutherford, who would receive the Nobel Prize in chemistry in 1908, remarked to his class at McGill University (Montreal, Canada) in 1898 that "... it is not safe or politic to invest much capital in a company for the transmission of signals by wireless [1]." The chief concerns of these early skeptics were limited range, reliability, and privacy. In reality, these concerns simply provided interesting challenges for future innovators.

Appleton and others throughout the first half of the twentieth century found through experimentation that when the carrier frequency of an electromagnetic wave was selected to match a particular channel, such as the ionosphere or the troposphere, surprisingly reliable worldwide communication was possible, depending on the particular time of day, season of the year, and sunspot activity [2]. Medium wave (100 kHz–3 MHz) and short wave (3 MHz–30 MHz) radiobands became the mainstay for the fledgling wireless broadcasting industry, as well as for ship-to-shore, telegraph, and military operations during the first several decades of the twentieth century, all relying on long-distance "skip" communications.

Amplitude modulation (AM) broadcasting, using medium wave frequencies in the 500 kHz–1700 kHz range, was launched throughout the world in the 1920s. Station KDKA, located in Pittsburgh, Pennsylvania, and owned by Westinghouse, was one of the first commercial AM broadcasting stations to go on the air in 1920. The initial broadcast provided listeners with the election results in which Warren G. Harding won the presidency over James Cox [3]. Inexpensive crystal radio sets were popular at the time and consisted of a small piece of germanium crystal that could be used with a conventional earpiece for local AM reception. An early (1922) factory

purchased crystals for 96 cents and from these developed radios that were sold for $2.25 [3]. To facilitate widespread adoption of AM reception by consumers, extremely low-cost receivers using envelope detectors were developed. Envelope detectors could be implemented easily with a simple diode and RC filter. The goal was to develop low-cost receivers in order to increase public access to this new form of communications.

Single-side band (SSB), which is a special form of AM, provides a spectrally efficient way of transmitting an AM waveform that also provides some security, because it cannot easily be detected by a standard AM envelope detector. In the early years of wireless communications, military personnel relied on SSB for wireless communications for both fixed and mobile applications throughout the world. Today, amateur radio operators and military operations (Military Amateur Radio Service — MARS) still use SSB because of its spectral efficiency.

In the 1920s and 1930s, Edwin Armstrong pioneered two fundamental inventions that still shape the wireless communications field [4]. As an engineer for the U.S. military, Armstrong invented the superheterodyne receiver. Using the concept of mixing, the superhetrodyne receiver (also called the superhet) allows a very high-frequency carrier wave to be translated down to a baseband frequency for detection. Until Armstrong's superhet design, receivers used direct conversion, where the receiver input signal was filtered directly from the antenna with a high-Q tunable bandpass filter, and then detected immediately. With the superhet receiver, it became possible to build much more sensitive receivers that could perform over a much wider frequency range, because it was no longer necessary to provide such a tight (and expensive) tunable filter at the incoming receiver frequency. Instead, a wider bandwidth fixed filter could be used at the antenna input, and a local oscillator could be tuned, thereby providing a mixed signal that could subsequently be filtered with better and less expensive filtering at a much lower intermediate (IF) frequency. The superheterodyne receiver allowed the received signal to be brought down to the baseband detector in stages. By standardizing on IF frequencies, component manufacturers were able develop devices such as filters, oscillators, and mixers that could be used over a wide range of wireless frequencies, thus allowing the wireless communications industry to begin its dramatic growth.

Armstrong's second invention, frequency modulation (FM), was patented in 1934 [3]. FM provided much greater fidelity than AM, as it was impervious to ignition and atmospheric noises that plagued AM transmissions. Since FM used constant envelope modulation, it also was much more power-efficient than AM, making it particularly well-suited for mobile radio telephone operation where battery preservation was key. The high-fidelity qualities of FM launched a new FM broadcasting industry, and frequency allocations for worldwide FM broadcasting were granted by the World Administrative Radio Conference (WARC) in the very high frequency (VHF) bands of 30 MHz–300 MHz, where wireless signals propagated reliably from a broadcast antenna to the visible horizon on earth. Unlike MF and HF waves, the shorter wavelengths of the VHF band are not generally propagated by skip mechanisms, thus providing much more predictable line-of-sight radio propagation behavior for both terrestrial and satellite use.[1] It is for this reason that virtually all modern wireless communication systems operate at or above the VHF frequency band.

Commercial television (TV) broadcasting evolved in the late 1940s and employed a type of AM modulation (vestigial sideband — VSB) for video transmission and FM for simultaneous aural transmission. Like FM broadcasting, TV broadcasting relied on the reliable "to-the-horizon" propagation offered by VHF radio waves as well as ultra-high-frequency (UHF) waves in the 300 MHz to 3 GHz bands.

Very early mobile telephone services also used FM for voice communications in the VHF and UHF bands, as mobile communications equipment became viable from a cost and reliability standpoint in the 1950s. Early taxicab and public service dispatch radio services soon realized that in order to handle a large population of mobile radio users with a finite set of channels, it would be necessary to employ trunking theory. Trunking theory determines how to allocate a finite set of channels to a large population of potential users, based on the calling patterns of an average user [6]. Early mobile radio systems used the concept of a control channel that is intermittently shared by all users of a radio service. The control channel is used by a central switch to broker instantaneous access to all of the available voice channels within the system. Trunking theory is used by a radio service to determine the appropriate number of channels to allocate for a large population of users, so that a particular average grade of service (or channel availability likelihood) can be provided to the users.

2. MOBILE CELLULAR TELEPHONY

The next major event in wireless communications was the development of cellular telephony. The development of cellular communications made mobile radio communications available to the general public at low cost. Cellular radio communications systems were developed in the United States by Bell Laboratories, Motorola, and other companies in the 1970s, and in Europe and Japan at about the same time. Test systems were installed in the United States in Washington, D.C., and Chicago in the late 1970s, and the first commercial cellular systems became operational in Japan in 1979, in Europe in 1981, and in the United States in 1983. Like other early cellular telephone systems, the first system in the United States used analog FM and operated in the UHF band. Designated AMPS (for advanced mobile phone system), it proved so successful that AMPS cellular telephones are still widely used today, especially in rural areas. The AMPS system is based on a channel spacing of 30 kHz.

[1] However, rare instances of skip from the ionosphere have been found to allow propagation distances of several thousand kilometers at frequencies well above 30 MHz [5].

In the early 1990s, the demand for cellular telephones exceeded available capacity, resulting in the development and adoption of so-called second-generation (2G) personal communications systems (PCS), with the first of these systems being fielded in the mid-1990s. All 2G PCS systems use the cellular concept, but employ digital transmission in place of analog FM. They have differing modulation and multiple access schemes as defined by their respective common air-interface standards. The European 2G standard, called global system for mobile communications (GSM), the Japanese system, and one U.S. standard [U.S. digital cellular (USDC) system] all employ time-division multiple access (TDMA), but with differing channel bandwidths and numbers of users per frame. A second U.S. 2G standard uses code division multiple access (CDMA). A goal of 2G system development in the United States was backward compatibility because of the large AMPS infrastructure that had been installed with the first generation. Europe, however, had several first-generation standards, depending on the country, and their goal with 2G was to have a common standard across all countries. As a result, GSM has been widely adopted, not only in Europe, but in much of the rest of the world. From the mid- to late 1990s, work began on third-generation (3G) standards, and these systems are beginning to be deployed after several widely publicized delays. A goal in the development of 3G systems is to have a common worldwide standard, but this proved to be too optimistic. Therefore, a family of standards was adopted, with one objective being to make migration from first-generation and second-generation systems possible.

Cellular systems were much more widely accepted by the public than was first expected when the first-generation systems were introduced. In many European and Pacific Rim countries, more than 50% of the population owns a cellular telephone, with the United States close behind. One might wonder why the United States is not leading the world in terms of cellular telephone usage. Perhaps there are three main reasons. The U.S. licensing process was not formalized until several years after the deployment of first-generation cellular systems in Europe and Japan. Also, the United States enjoyed the preexistence of a very good wireline system before cellular telephones were introduced. Finally, the United States had the practice of charging the cellular subscriber for both incoming and outgoing calls. It appears that the billing mechanism is changing with the further expansion of cellular telephone use in the United States, with cellular service providers not only trying to attract new customers but also trying to reduce the "churn," or changing of the customer from one provider to another.

2.1. Basic Principles of Cellular Radio[2]

Radio telephone systems had been in use long before the introduction of cellular radio, but their capacity was very limited because they were designed around the concept of a single base station servicing a large area — often the size of a large metropolitan area. Cellular telephone systems are based on the concept of dividing the geographic service area into a number of cells and servicing the area with low-power base stations placed within each cell, usually the geographic center. This allows the band of frequencies allocated for cellular radio use (currently there are two bands in the 900 and 1800 MHz regions of the radio spectrum) to be reused in physically separated geographic regions, depending on the accessing scheme involved. For example, with AMPS, the same radio channels are reused over a relatively small geographic area and are repeated once every seven cells. In today's CDMA (code division multiple access), the same channels are used in each cell. Another characteristic that the successful implementation of cellular radio depends on is the attenuation of transmitted power with frequency. For free space, power density decreases per the inverse square of the distance from the transmitter. However, because of the characteristics of terrestrial radio propagation, the decrease of power with distance is greater than an inverse square law, typically between the inverse third and fourth power of the distance. Were this not the case, the cellular concept would not work. Since the service area (the geographic area of interest to be covered with cellular service) is represented by tessallating cells overlaid on a map of the coverage region, it is necessary for the mobile user to be transferred from one base station to another as the mobile moves within the service area. This procedure is called *handoff* or *handover*. Also note that it is necessary to have some way of initializing a call to a given mobile and keeping track of it as it moves from one base station to another. This is the function of a *mobile switching center* (MSC). MSCs also interface with the public switched telephone network (PSTN).

Consider Fig. 1, which shows a typical cellular tessellation using hexagons to represent cells. It is emphasized that real cells are never hexagonal; indeed, some cells may have very irregular shapes because of

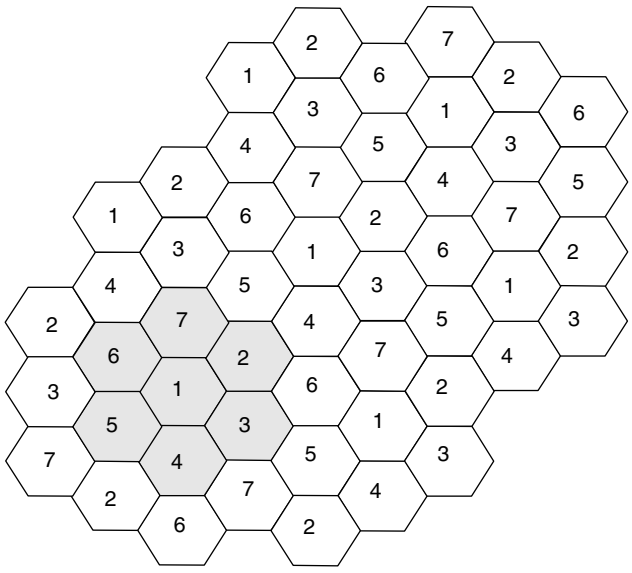


Figure 1. Hexagonal grid system representing cells in a cellular radio system; a reuse pattern of seven is illustrated.

[2] This section follows closely a previous publication by Ziemer and Tranter [7].

geographic features and illumination patterns of the transmit antenna. However, hexagons are typically used in theoretical discussions of cellular radio because a hexagon is one geometric shape that tessellates a plane and very closely approximates a circle, which is what we assume for the contours of equal transmit power in a relatively flat environment. Note that a seven-cell reuse pattern is indicated in Fig. 1 via the integers given in each cell. Obviously, there are only certain integers that work for reuse patterns, for example, 1, 3, 4, 7, 9, 12, ... A convenient way to describe the frequency reuse pattern of an ideal hexagonal tessellation is to use a nonorthogonal set of axes, U and V, intersecting at 60 degrees as shown in Fig. 2. The normalized grid spacing of one unit represents the distance between adjacent base stations, or hexagon centers. Thus, each hexagon center is at point (u, v) where u and v are integers. Using this normalized scale, each hexagon vertex is $R = 1/\sqrt{3}$ from the hexagon center. It can be shown that the number of cells in an allowed frequency reuse pattern is given by

$$N = i^2 + ij + j^2 \tag{1}$$

where i and j take on integer values. Letting $i = 1$ and $j = 2$ (or vice versa), it is seen that $N = 7$, as we already know from the pattern identified in Fig. 2. Considering other integers, the number of cells in various reuse patterns are as given in Table 1. Currently used reuse patterns are 1 (CDMA), 3 (GSM), and 7 (AMPS).

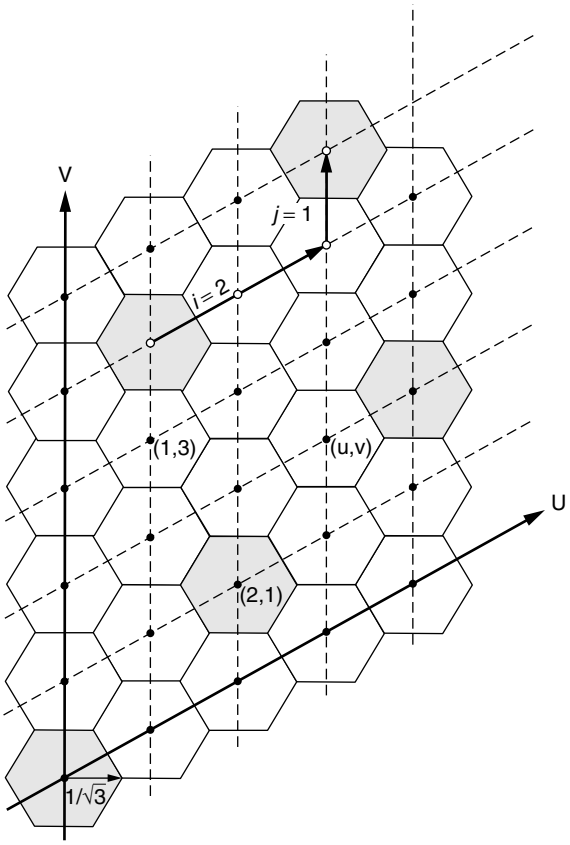


Figure 2. Hexagonal grid geometry showing coordinate directions; a reuse pattern of seven is illustrated.

Table 1. Possible Reuse Patterns in Cellular Systems

Reuse Coordinates		Number of cells in Reuse Pattern	Normalized Distance Between Repeat Cells
i	j	N	$\sqrt{N}$
1	**0**	1	1
2	**1**	3	1.732
1	**2**	7	2.646
2	**2**	12	3.464
1	**3**	13	3.606
2	**3**	19	4.359
1	**4**	21	4.583
2	**4**	28	5.292
1	**5**	31	5.568

Another useful relationship is the distance between like-cell centers, D_{co}, which can be shown to be $D_{co} = \sqrt{3N}R$, which is $D_{co} = \sqrt{N}$ since $R = 1/\sqrt{3}$. This is an important consideration in computing *cochannel interference*, that is, the interference from a second user in a nearby cell that is using the same frequency assignment as that of a user of interest. Clearly, if a reuse pattern has N cells in it, this interference could be a factor of N larger than that due to a single interfering user (not all cells at distance $\sqrt{N}$ from a user of interest may have an active call on that particular frequency). Note that there is a second ring of cells at $2\sqrt{N}$ that can interfere with a user of interest, but these usually are considered to be negligible compared with those within the first ring of interfering cells.

Assume a decrease in power with distance, R, of the form

$$P_r(R) = K\left(\frac{R_0}{R}\right)^{\alpha} \text{ watts} \tag{2}$$

where R_0 is a reference distance and the power is known to be K watts. As mentioned previously, the power law is typically in the range of 2.5 to 4 for terrestrial propagation, which can be analytically shown to be a direct consequence of the earth's surface acting as a partially conducting reflector (other factors such as scattering from buildings and other large objects also come into play, which accounts for the variation in α). In logarithmic terms, the received power is

$$P_{r,\text{dBW}}(R) = K_{\text{dB}} + 10\alpha \log_{10} R_0 - 10\alpha \log_{10} R \text{ dBW} \tag{3}$$

Now consider reception by a mobile from a base station of interest, A, at distance d_A, while at the same time being interfered with from a cochannel base station B, at distance D_{co} from A. We assume for simplicity that the mobile is on a line connecting A and B. Thus, the signal-to-interference ratio (SIR) in decibels is

$$\text{SIR}_{\text{dB}} = K_{\text{dB}} + 10\alpha \log_{10} R_0 - 10\alpha \log_{10} d_A - [K_{\text{dB}} + 10\alpha \log_{10} R_0 - 10\alpha \log_{10}(D_{co} - d_A)] \tag{4}$$

This gives

$$\mathrm{SIR_{dB}} = 10\alpha \log_{10}\left(\frac{D_{co}}{d_A} - 1\right) \mathrm{dB} \tag{5}$$

Clearly, as $d_A \to D_{co}/2$, the argument of the logarithm approaches 1 and the SIR_{dB} approaches 0. As $d_A \to D_{co}/2$, the mobile should ideally hand off from A and begin using B as its base station.

We can also compute a worst-case SIR for a mobile of interest. If the mobile is using base station A as its source, the interference from the six other cochannel base stations in the reuse pattern is no worse than that from B (the mobile is assumed to be on a line connecting A and B). Thus, $\mathrm{SIR_{dB}}$ is underbounded by

$$\begin{aligned}\mathrm{SIR_{dB,min}} &= 10\alpha \log_{10}\left(\frac{D_{co}}{d_A} - 1\right) - 10\log_{10}(6)\ \mathrm{dB} \\ &= 10\alpha \log_{10}\left(\frac{D_{co}}{d_A} - 1\right) - 7.7815\ \mathrm{dB}\end{aligned} \tag{6}$$

2.2. The Mobile Wireless Channel

A distinguishing factor in terrestrial mobile radio is the channel impairments experienced by the signal. In addition to the Gaussian noise present in every communication link due to the nonzero temperature of the receiver, and the cochannel interference, another important source of degradation is the mobile wireless channel. As the mobile moves, the signal strength varies drastically because of multiple transmission paths as well as objects that block the line-of-sight propagation path. The mobile wireless channel induces the attenuation and distortion seen at the receiver due to the environment and mobility. The attenuation and distortion caused by the mobile wireless channel can be broken into two main components: large-scale fading and small-scale fading.

2.2.1. Large-Scale Fading. Large-scale fading is related to the path loss discussed earlier. As mentioned previously, the terrestrial wireless channel (because it typically involves scattering and possibly non–line-of-sight conditions) experiences a loss in received power that increases with distance raised to the third to fourth power. This typically is called the path loss exponent. Additionally, for a given distance there is some statistical variation about the distance-dependent mean value represented by the path loss exponent. This variation normally is attributed to large objects in the environment and is termed *shadowing*. For a fixed propagation distance R, some radial paths from a transmitter experience more shadowing than others as a result of the spatial variations of objects over a particular geometry. The large-scale variation about a distance-dependent mean signal level typically follows a log-normal distribution with a standard deviation of 8–12 dB [6].

2.2.2. Small-Scale Fading. Small-scale fading refers to variations in the signal strength seen by a mobile receiver as it moves over very short (on the order of wavelengths) distances. This type of fading can be characterized in terms of a Doppler spectrum, which is determined by the motion of the mobile (and to a small degree the motion of the surroundings, such as a wind blowing trees or the motion of reflecting vehicles). Another characteristic of small-scale fading is time-varying delay spread due to the differing propagation distances of the received multipath components. As signaling rates increase, this becomes a more serious source of degradation due to intersymbol interference (ISI) of the transmitted signal. Equalization can be used to compensate for ISI.

Small-scale multipath can be divided into two types: *unresolvable*, where the resultant at the receiver can be approximated as a sum of phasors whose amplitudes and phases vary with motion of the transmitter, receiver, or environment; and *resolvable*, where propagation times are long compared with the inverse signal (receiver) bandwidth. The latter condition means that the channel is frequency selective (the signal bandwidth is wide with respect to variations in channel frequency response). It should be clear that multipath resolvability depends on receiver bandwidth relative to the time intervals between multipath components.

A common technique used to combat small-scale fading is diversity. In GSM and USDC, diversity takes the form of error-correction coding using interleaving. For CDMA, diversity can be added in the form of simultaneous reception from two different base stations near cell boundaries (soft handoff). Other combinations of simultaneous transmissions [9] and receptions in a rich multipath environment are being proposed for future-generation systems to significantly increase capacity (this is also called transmit diversity). Also used in CDMA is a tool called a RAKE receiver, which detects the separate resolvable multipath components and puts them back together in a constructive fashion. In general, the various diversity techniques can be divided into three general categories:

1. Space diversity — the use of more than one antenna to capture the propagating signal.
2. Frequency diversity — the use of several carrier frequencies or the use of a wideband signal with an equalizer or RAKE receiver
3. Time diversity — spreading out the effects of errors through interleaving and coding

All three of these techniques are commonly used.

2.3. Multiple Access Techniques

The implementation of a cellular radio system depends heavily on the use of a multiple access scheme. Multiple access is the technique that allows multiple users to access the system simultaneously. We have already touched on the idea of multiple access, but in this section we describe it in more detail. Historically, three common methods for multiple access are

1. Frequency division multiple access (FDMA)
2. Time division multiple access (TDMA)
3. Code division multiple access (CDMA)

Figure 3 schematically illustrates these three access schemes. Three dimensions are shown in each figure — time, frequency, and code. In Fig. 3 (a), this time-frequency-code resource is split into a number of frequency channels, each one of which may be assigned to a different user. In other words, we give potential users access to the communication resource using FDMA. In Fig. 3 (b), the time-frequency-code resource is split into a number of time slots, each of which may be assigned to a different user. In other words, we give potential users access to the communication resource using TDMA. Finally, in Fig. 3 (c), the time-frequency-code resource consists of a number of codes, each of which may be assigned to a different user. In other words, we give potential users access to the communication resource using CDMA.

In cellular radio systems, two of these typically are used together. For example, in the global system for mobile (GSM) communications system, TDMA and FDMA are used together in that the allocated frequency spectrum is divided into 200-kHz chunks for each TDMA frame, and each frame can then accommodate up to eight users using TDMA. As another example, the IS-95 standard was designed around the use of CDMA to accommodate up to 61 users (There are 64 potential channels. Three channels are set aside for synchronization, the pilot, and for paging. See Ref. 8). A Walsh code is assigned to each of the 64 channels, and each block of 61 users requires only 1.25 MHz of spectrum. Thus, the allocated frequency spectrum is divided into multiple 1.25-MHz chunks of frequency, each of which can be employed for up to 61 users that are distinguished from each other by their assigned CDMA codes.

3. CHARACTERISTICS OF SECOND-GENERATION CELLULAR SYSTEMS

As mentioned previously, the development of cellular telephony is commonly broken up into three distinct stages termed generations. First-generation cellular systems (including the AMPS standard in the United States) were analog systems that carried only voice traffic. As the capacity of first-generation systems filled, more efficient cellular systems were developed. These so-called second-generation systems used digital modulation and allowed roughly a three times capacity improvement over the analog systems. Also, because they were digital, these second-generation systems were capable of carrying rudimentary data services.

Space does not allow much more than a cursory glance at the technical characteristics of the most popular second-generation cellular radio systems — in particular, USDC, GSM, and CDMA (referred to as IS-95 in the past, where the "IS" stands for "interim standard"). For complete details, the standard for each may be consulted. Before doing so, however, the reader is warned that this amounts to thousands of pages in each case. Table 2 summarizes

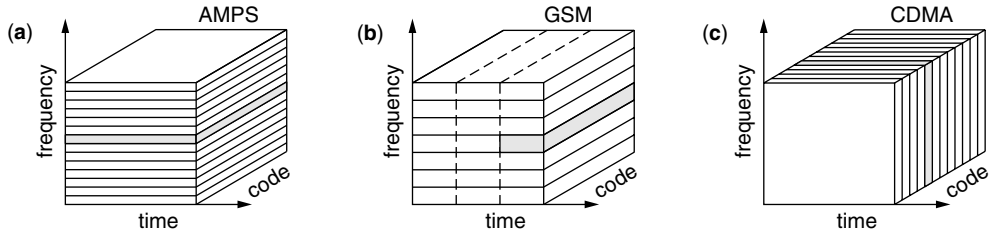


Figure 3. Illustrations of multiple access schemes: (**a**) FDMA; (**b**) TDMA; (**c**) CDMA.

Table 2. 3G AMPS, GSM, and CDMA Technologies Compared

	AMPS	GSM	CDMA
Carrier separation	30 kHz	200 kHz	1.25 MHz
No. channels/carrier	1	8	61
Accessing techniques	FDMA	TDMA-FDMA	CDMA-FDMA
Frame duration	NA	4.6 ms with 0.58 ms slots	20 ms
User modulation	FM	GMSK, $BT = 0.3$ Binary, diff. encoded	BPSK, FL 64-ary orthog. RL
Cell reuse pattern	7	3	1
Cochan. Inter. Protect.	$\leq$15 dB	$\leq$12 dB	NA
Error correction-coding	NA	Rate 1/2 convolutional Constraint length 5	Rate 1/2 convol., FL Rate 1/3 convol., RL Both constr. length 9
Diversity methods	NA	Freq. hop, 216.7 hops/s Equalization	Wideband signal Interleaving RAKE
Speech representation	Analog	Residual pulse excited, Linear prediction coder	Code-excited vocoder
Speech coder rate	NA	13 kbps	9.6 kbps max

some of the most pertinent features of these three systems. For further details, see Rappaport [6] and Ziemer and Peterson [8].

4. THIRD-GENERATION CELLULAR RADIO

Second-generation systems provides one of the most successful practical applications of many aspects of communication theory, including speech coding, modulation, channel coding, diversity techniques, equalization, and so on. The implementation of 3G cellular promises the same accommodation as 2G for voice in addition to much higher data rate capacity, including 64 kbps for high-speed vehicles over wide areas, 384 kbps for pedestrian speeds over smaller areas, and 2 Mbps for stationary (but movable) locations over building- or campus-size areas. With these capabilities, 3G cellular indeed promises ready access to information anytime and anywhere. Specifically, 3G will include the following options:

1. Flexible support of multiple services (data rates from kbps to Mbps; packet transmission)
2. Voice
3. Messaging — email, fax, etc.
4. Medium-rate multimedia — Internet access, educational
5. High-rate multimedia — file transfer, video
6. High-rate interactive multimedia — video teleconferencing, telemedicine, etc.
7. Mobility: quasi-stationary to high-speed platforms
8. Global roaming: ubiquitous, seamless coverage
9. Evolution from second-generation systems

Two CDMA-based radio transmission technologies (RTTs) have emerged as the leading answers for 3G radio. One, called WCDMA for wideband CDMA, originated out of several European and Japanese studies carried out in the mid-1990s. The other, called CDMA2000, originated from a coalition of U.S. companies working out a standard in the late 1990s that would provide an easy migration path for the CDMA 2G standard, IS-95. The characteristics of CDMA and CDMA2000 are summarized in Table 3.

Table 3. 3G Radio Transmission Technologies Compared

PARAMETER	WCDMA	CDMA2000
Carrier spacing	5 MHz	3.75 MHz
Chip rate	3.84 Mcps	3.684 Mcps
Data modulation	BPSK	DL — QPSK; UL — BPSK
Spreading	OQPSK	OQPSK
Power control frequency	1500 Hz	800 Hz
Variable data rate implementation	Variable SF; multicode	Repetition, puncturing, multicode
Frame duration	10 ms	20 ms
Coding	Turbo and convolution	Turbo and convolution
Base station synchronized?	Asynchronous	Synchronous
Base station acquisition/detect	3 step: slot, frame, code	Time shifted PN correl.
Forward link pilot	TDM dedicated pilot	CDM common pilot
Antenna beam forming	TDM dedicated pilot	Auxiliary pilot

Table 4. Rate Examples for the Uplink of WCDMA

Number of Data Channels	Bits per Slot	Spreading Factor	Channel Symbol Rate (kbps)	Data Rate (kbps)
1	10	256	15	7.5
1	20	128	30	15
1	40	64	60	30
1	80	32	120	60
1	160	16	240	120
1	320	8	480	240
1	640	4	960	480
6	640	4	5740	2370

5. CONCLUSIONS

Despite the fact that wireless communications is over 100 years old, it remains an active area of research, development, and commercialization. Cellular systems continue to grow in popularity, with penetration rates increasing all over the world. Wireless networks are becoming more popular in business and campus environments. Wireless mobile radio systems provide the ability to communicate and access information from any location. As the need or desire for these increases in both the business world and our daily lives, the need for seamless connectivity will grow. As a result, it is expected that mobile radio systems will continue to be an important technology for many years to come. The interested reader is encouraged to investigate this topic further. Refs. 6 and 8 provide a starting point for this endeavor.

BIBLIOGRAPHY

1. J. Campbell, *Rutherford: Scientist Supreme*, Christchurch, NZ, AAS Publications, 1999.
2. E. V. Appleton and W. J. G. Beynon, The application of ionospheric data to radio communication problems: Part I, *Proc. Phys. Soc.* **52**: 518–533 (1940).
3. T. Lewis, *Empire of the Air: The Men Who Made Radio*, New York, HarperCollins, 1991.
4. *A History of the Radio Club of America, Inc.: 1909–1984, Seventy-Fifth Diamond Jubilee Yearbook*, Radio Club of America, Inc., 1984. (Library of Congress Catalog Number 84–061879)
5. T. S. Rappaport, R. L. Campbell, and E. Pocok, A single-hop F2 propagation model for frequencies above 30 MHz and

path distances greater than 4000 km, *IEEE Trans. Antennas Propag.* **38**: 1967–1968 (1990).

6. T. S. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed., Prentice Hall PTR, Upper Saddle River, NJ, 2002.
7. R. E. Ziemer and W. H. Tranter, *Principles of Communications: Systems, Modulation and Noise*, 5th ed., Wiley, New York, 2002.
8. R. E. Ziemer and R. L. Peterson, *Introduction to Digital Communications*, 2nd ed., Prentice Hall PTR, Upper Saddle River, NJ, 2001.
9. S. M. Alamouti, A simple transmit diversity technique for wireless communications, *IEEE J. Select. Areas Commun.* **16**: 1451–1458 (1998).

MODELING AND ANALYSIS OF DIGITAL OPTICAL COMMUNICATIONS SYSTEMS

MARK SHTAIF
Tel-Aviv University
Tel-Aviv, Israel

1. INTRODUCTION

Since the early 1990s the world of optical communications has experienced staggering growth driven by an unprecedented acceleration of demand. Technologically, what enabled this growth was major advancements in the area of fiberoptic components, the most significant of which were the invention of the erbium-doped fiber amplifier (EDFA) [1,2] and the implementation of devices that compensate for the chromatic dispersion of optical fibers [3]. These technologies revolutionized almost every aspect of fiberoptic transmission. They allowed systems to extend to multiple thousands of kilometers without electronic regeneration and made the concept of wavelength-division multiplexing (WDM) cost-effective for the first time, thereby increasing the aggregate capacity of optical systems by many orders of magnitude. Simultaneously with the enhancement in performance, the combination of optical amplifiers with effective dispersion compensation technology has also changed the way in which fiberoptic systems operate. Systems are no longer limited by the optical signal power impinging on the photoreceiver, and therefore effects such as shot noise and even thermal noise generated in the receivers became practically irrelevant. Instead, the dominant noise source is amplified spontaneous emission generated in the amplifiers themselves. Waveform distortions are no longer dominated by the chromatic dispersion of the link; instead it is the optical nonlinearity of the transmission fiber that has become the chief source of distortions and interference. Now, more than ever before, the performance of optical communications systems is dictated not by the imperfection of individual components but chiefly by fundamental physical principles pertaining to signal generation, transmission, and detection. The description of those principles is the main goal of this article.

The structure of a typical optical communication system is illustrated in Fig. 1. It consists of a stack of transmitters, a fiberoptic link, and a stack of optical receivers. The light generated by each transmitter has a specific central optical wavelength (or frequency), and the signals emitted by the various transmitters are optically multiplexed into a single fiber. On the receiver side the optical channels are demultiplexed such that each channel is fed into its own dedicated receiver. The link consists of multiple spans of fiber separated by optical amplifiers. The typical length of a single fiber span ranges from 40 to 120 km, and the length of the entire link varies between a few tens of kilometers (single span) in short-reach applications and 9000 km in transpacific systems. Each amplifier provides gain to compensate for the attenuation incurred in the preceding fiber span, and in most cases it also incorporates a dispersion-compensating module (DCM) that is included in its physical structure and whose role it is to balance the chromatic dispersion in the transmission fiber. As is always the case when modeling complicated systems, certain assumptions need to be made regarding its various components. In our case, since we wish to focus on fundamental limitations, we shall ignore the detailed description of components that do not impose fundamental constraints on system performance. Thus we shall assume that the optical transmitter is capable of generating any reasonable pulseshape that we desire, an assumption that is consistent with most waveforms that are considered favorable for fiberoptic transmission. We will also assume that the modulated electric field of each channel can be approximated as $\sum_k a_k g(t - kT) \exp(-i\omega_{jt})$, where a_k represents the value of the kth symbol, $g(t)$ is the slowly varying envelope of an individual pulse, and ω_j is the central optical frequency of the jth channel. The multiplexer and the demultiplexer used on the two sides of the system are assumed to be perfect in the sense that they combine and separate the individual channels, respectively, without affecting their waveforms. The photodetector is described as a module that generates an electric current that is proportional to the incident optical power. Finally, noise mechanisms such as shot noise, laser noise, and thermal noise at the receiver are ignored as they are negligible relative to the noise contributed by amplified spontaneous emission.

The dominant mechanisms that determine the performance of optical systems are the noise generated in the amplifiers on one hand and waveform distortions taking place in the optical fiber on the other hand. The distortions are due to a combination of fiber dispersion, nonlinearity, and polarization related effects. Although, as pointed out earlier, dispersion by itself can be perfectly compensated for, it affects the way in which fiber nonlinearities distort the optical waveforms, and therefore the interplay between the two is of outmost importance. Polarization-related distortions are caused by mechanical stress and geometric imperfections that break the cylindrical symmetry of optical fibers, thereby making its transmission properties polarization-dependent. They are particularly important in systems operating with exceptionally bad fiber [4], or when the data rates per channel are particularly high [$\geq$40 Gbps (gigabits per second)]. In the following section we discuss the modeling of optical amplifiers. Section 3 deals with the modeling of receivers and

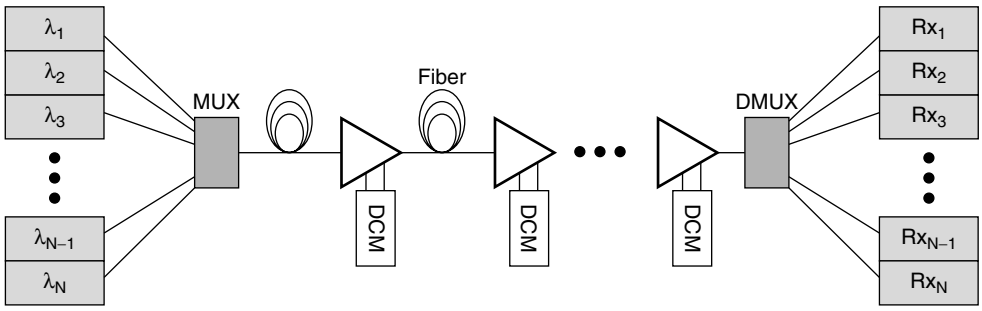


Figure 1. A schematic description of a WDM optical system. The transmitters are labeled by their central wavelengths λ_1 to λ_N. Triangles represent optical amplifiers, and DCM stands for dispersion compensation module.

the performance of systems limited by optical noise. The issues of fiber transmission are dealt with in Sections 4 and 5; Section 4 concentrates on the combination of chromatic dispersion and nonlinearities, and Section 5 reviews polarization-related effects.

2. AMPLIFICATION AND NOISE

One of the most fundamental properties of coherent optical amplification is that it is always accompanied by noise [5–7]. The principles that necessitate this noise and determine its properties can be fully understood only in a quantum-mechanical context, but in all relevant situations when the number of noise photons emitted by the amplifier is much greater than 1, the effect of an optical amplifier has a very accurate classical representation [8]

$$\vec{E}_{\text{out}} = \sqrt{G}\vec{E}_{\text{in}} + \vec{n}(t) \tag{1}$$

where $\vec{E}_{\text{in}}$ and $\vec{E}_{\text{out}}$ are the input and output electric field vectors, respectively, G is the power gain of the amplifier, and $\vec{n}(t)$ is a white Gaussian noise process whose power density spectrum measured along any given state of polarization is $\hbar\omega_0 n_{\text{sp}}(G-1)$. Here ω_0 is the central frequency of the amplified signal, $\hbar$ is Planck's constant, and the term n_{sp}, which is always greater than or equal to 1, accounts for the enhancement of noise in amplifiers that contain loss mechanisms. In amplifiers based on the inversion of carrier populations, losses may result from incomplete inversion, and therefore n_{sp} is commonly known as the inversion factor. Although physically it is the gain of the amplifier and its inversion factor that determine the noise power, amplifier manufacturers and designers often talk about optical amplifiers in terms of their noise figure (NF). The concept of the NF is borrowed from RF amplification, and it describes the deterioration in the signal-to-noise ratio (SNR) of a coherent (shot-noise-limited) signal as a result of optical amplification [9]. Unlike its RF equivalent, the optical noise figure definition is based on the SNR obtained in a measurement of optical energy, where the noise is not additive and therefore the interpretation of the NF is not obvious. A useful relation illustrating the significance of the noise figure can be obtained in the case of high-gain amplifiers, where it can be shown that $\text{NF} \simeq 2n_{\text{sp}}$.

As depicted in Fig. 1, optical systems usually consist of a large number of amplified spans such that each amplifier compensates for the loss of the fiber span that precedes it. In principle, the length of an individual span is a free parameter. One may choose to implement a system with a large number of short spans or a small number of long spans such that in both cases the signal power impinging on the receiver is identical. Yet when it comes to the accumulation of noise, the two scenarios are considerably different from each other. To observe the difference, we express the total noise power at the receiver within a signal bandwidth B as $P_{\text{ASE}} = \hbar\omega_0 n_{\text{sp}}(G-1)BN$, where N is the number of spans. If we denote the overall length of the system by L and the loss coefficient of the fiber by α, then the amplifier gain can be expressed as $G = \exp(\alpha L/N)$ such that it exactly compensates for the losses of the span, and therefore we obtain $P_{\text{ASE}} = \hbar\omega_0 n_{\text{sp}}\alpha LB(G-1)/\ln(G)$. Notice that the minimum noise power is equal to $P_{\text{ASE}} = \hbar\omega_0 n_{\text{sp}}\alpha LB$ and corresponds to the case where the number of spans approaches infinity or $G \to 1$. The penalty for introducing a finite number of spans is described by the function $(G-1)/\ln(G)$, which is plotted in Fig. 2. Notice that for typical gain values in optical systems (of the order of 20 dB) the dependence of the penalty on G is almost linear. The preceding calculation of the noise power assumed that the power that is launched into the system is independent of G, which implies that as we reduce G, the signal power averaged over the system length increases. This can be easily visualized in the limit in which the number of spans goes to infinity and G approaches 1, where the path-averaged power becomes equal to the power launched into the system. As we shall see in the next section, an increase in the optical power along the system enhances the nonlinear effects and increases waveform distortions. Therefore, a more meaningful quantity than the noise power calculated above is the ratio between the signal and noise powers, assuming that the launched signal power is adjusted as a function of G such that the path-averaged power is kept constant [10]. In this case the penalty for using a finite number of spans is given by $F(G) = [(G-1)/\ln(G)]^2/G$ and is shown by the dashed curve in Fig. 2. The obvious advantage in reducing the length of the span and increasing the number of amplified spans in an actual system is balanced by the higher cost of systems containing a larger number of amplifiers.

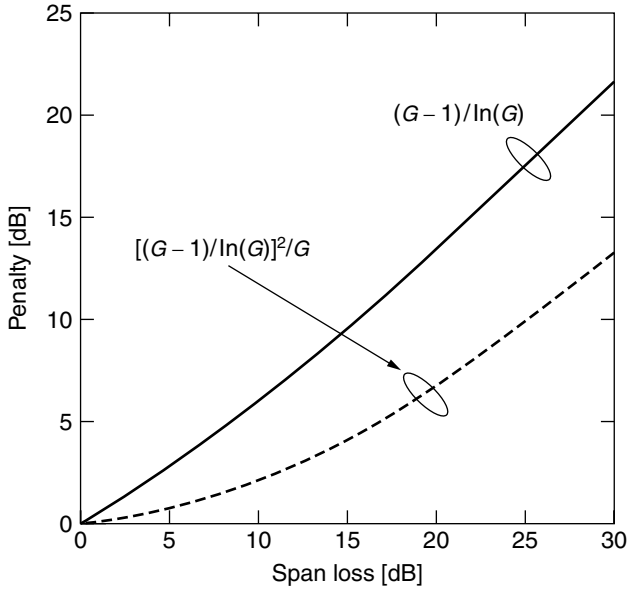


Figure 2. The optical SNR penalty caused by the use of lumped amplification as a function of the gain G (which is equal to the span losses). The solid curve corresponds to the case of fixed launched power, and the dashed curve represents the case of fixed path-averaged power. (After Ref. 10.)

3. MODELING OF OPTICAL RECEIVERS

The detection of light is based on the excitation of electrons by photons in photosensitive materials. This process generates an electric current whose average is proportional to the power of the incident light. The fluctuations around this average are caused by shot noise and as we have noted earlier, they are negligible in amplified systems where the dominant noise contribution comes from spontaneous emission. The vast majority of receivers that are used for optical communications are based on the direct detection of light and are therefore sensitive only to the incident optical power. Receivers that are also capable of detecting the optical phase are called *coherent receivers*, and their principle of operation relies on the coupling of the incoming optical field with a strong local oscillator prior to photodetection [11]. Coherent receivers offer two major advantages: (1) both the intensity and the phase of the optical field can be used for transmitting information and (2) they provide "free" amplification since the measured signal is proportional to the product of the incident optical field with an arbitrarily intense local oscillator. Historically, it was primarily the second advantage that drove the entire field to work on coherent transmission, and therefore with the invention of efficient optical amplifiers, interest in this topic became marginal. But the more fundamental reason for the loss of interest in coherent optical systems is provided by Gordon and Mollenauer [12] and is related to the fact that the optical phase is very easily corrupted by the combined effect of noise and fiber nonlinearities. Therefore in systems that are not limited by the available optical signal power, higher capacities can be achieved when only intensity modulation is used.

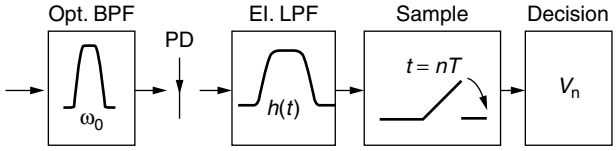


Figure 3. A schematic description of an intensity receiver (Opt. BPF — optical bandpass filter, PD — photodiode, El. LPF — electrical lowpass filter).

The structure of a generic intensity receiver is illustrated in Fig. 3. The optical signal is first filtered optically and photodetected. The electric current generated by the photodetector is filtered electrically and sampled. The samples are then fed into a decision circuit that determines the identity of the transmitted symbol. The sampling period and phase are matched to the received data by a separate clock recovery mechanism that is not shown in the figure. The modulation format that is almost exclusively used in intensity modulated systems is on/off keying, where logical ones and zeros are indicated by the existence or the absence of an optical pulse, respectively. The received signal in this case can be expressed by $\sum_k a_{kg}(t - kT)\exp(-i\omega_{jt})$ where the value of a_k is either 0 or 1. In spite of numerous attempts, modulation schemes using more than two intensity levels have not yet proved themselves usable in optical communications, as we shall explain later in this section. The details of the electric filter in the receiver and its model vary from system to system. Nevertheless, valuable insight into the problem can be obtained by assuming that the impulse response of this filter is square; that is, it is equal to 1 in the time interval between 0 and T and to zero otherwise. Then the effect of the filter is simply to integrate the optical energy of the received signal within the symbol duration so that the value of the kth sample is given by

$$V_k = \int_0^T dt|a_k|^2|g(t) + n(t)|^2 \tag{2}$$

where we have neglected the effect of intersymbol interference (ISI). This simplified version of the optical receiver is commonly referred to as "integrate and dump" and it is attractive as it allows convenient analytic handling. In particular, it can be shown [13] that the probability distributions of V_k corresponding to $a_k = 0$ and $a_k = 1$ are the central and the noncentral chi-square distributions, respectively, with mBT degrees of freedom [14]. The coefficient m is equal to 2 when a properly aligned polarizer is used to select the polarization of the received signal prior to photodetection, whereas in the absence of a polarizer $m = 4$. Figure 4a shows the two probability density functions corresponding to the case of $BT = 5$, $m = 4$, and where the ratio between the optical signal and noise powers (evaluated in a bandwidth B) is equal to 4. This would be the case for example in a typical 10-Gbps transmission system using a 50-GHz optical filter and characterized by an optical SNR (measured in a bandwidth of 0.1 nm) of approximately 12 dB. The same curves are also shown on a logarithmic scale in Fig. 4b. The

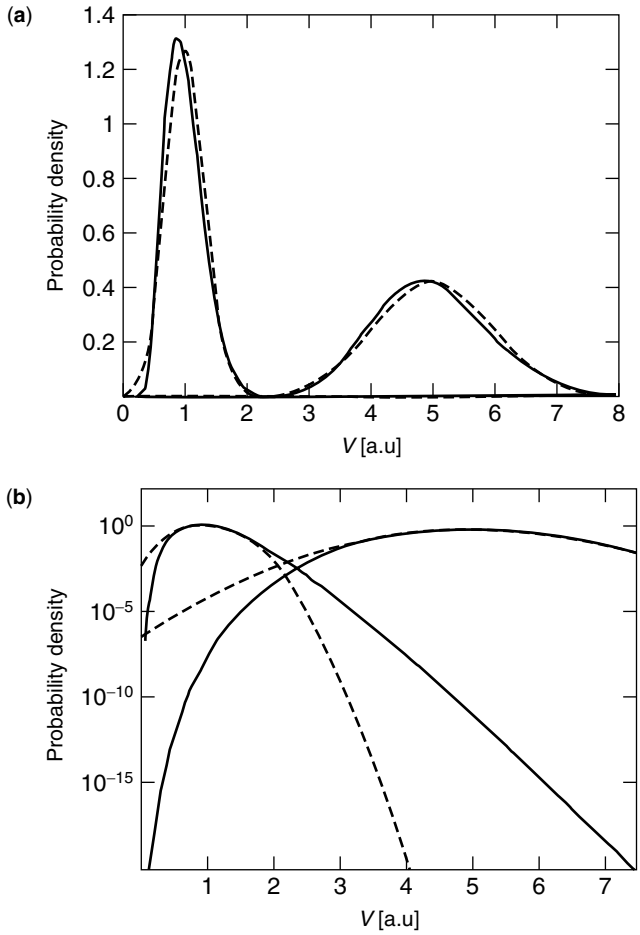


Figure 4. The probability density functions corresponding to the transmission of a logical 0 and 1: (**a**) linear scale; (**b**) logarithmic scale. The solid curves are the exact chi-square distributions, and the dashed curves represent a Gaussian fit to these distributions.

optimal decision threshold is equal to the value of V at the point where the two distributions intersect. Interestingly, the error probabilities given that the transmitted symbol is either 0 or 1 can be shown to be almost identical in a broad range of system parameters, so that the optical channel can be very accurately approximated as symmetric. Figure 4 also shows for comparison the Gaussian distribution functions (dashed curves) that are obtained based on the mean and the variance of the sampled signal. Although the Gaussian and the chi-square distributions are visibly different, it is common practice among optical system engineers to estimate system performance according to the assumption that the received signals are Gaussian distributed. Coincidently, in spite of the apparent difference between them, the two kinds of distributions give very similar average error rates [13]. This fact is illustrated in Fig. 5, which shows the ratio between the approximate Q factor resulting from the Gaussian approximation and the actual Q factor over a broad range of values [where the Q factor is related to the error probability in the usual way: $p_{\text{err}} = 1/\sqrt{2\pi}\int_Q^\infty \exp(-x^2/2)\,dx$]. While Fig. 5 justifies the

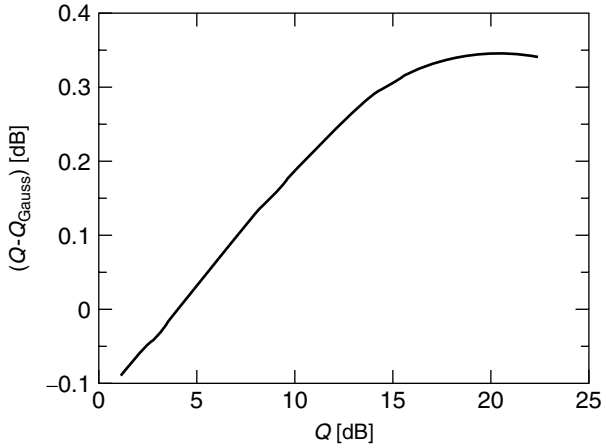


Figure 5. An illustration of the accuracy of the Gaussian approximation in estimations of the average error rate. The vertical axis (ordinate) shows the difference between the Q factor obtained from the Gaussian approximation and the actual Q factor obtained with the chi-square distributions. The horizontal axis corresponds to the actual Q.

use of the Gaussian approximation, it is important to emphasize that it is only the average error rate that can be accurately evaluated this way. The threshold level and the individual error rates conditioned on the transmission of either 0 or 1 cannot be obtained correctly from the Gaussian approximation.

To conclude the subject of linear transmission, we address the question of the ultimate limit to the information capacity of an optical channel. This question can be easily answered in the case of coherent optical systems where Shannon's capacity formula [15] corresponding to channels affected by additive Gaussian noise can be directly applied:

$$C_{\text{coh}} = 4 \times \frac{1}{2T}\log_2\left(1 + \frac{\mathcal{E}}{N_0}\right) \tag{3}$$

where $\mathcal{E}$ is the average energy per symbol duration and N_0 is the power density of the noise measured on both quadratures and both polarizations. The factor of 4 in front of the expression is due to the fact that the signal can be transmitted over 4 degrees of freedom (i.e., two states of polarization and two orthogonal quadratures). In the case of an intensity-modulated system the general calculation of the capacity is very difficult, but a simple and intuitive solution can be obtained in the limit where the optical SNR is much greater than 1, as is almost always the case in actual systems. In this limit it can be shown that the channel capacity (for both states of polarization) is given by [16]

$$C_{\text{Int}} \simeq \frac{1}{T}\log_2\left(\frac{\mathcal{E}}{2N_0}\right) \tag{4}$$

As usual, the capacity (4) represents the highest transmission rate in bits per second that can be reliably achieved in an intensity-modulated optical channel. It implies optimal coding and involves no constraints on

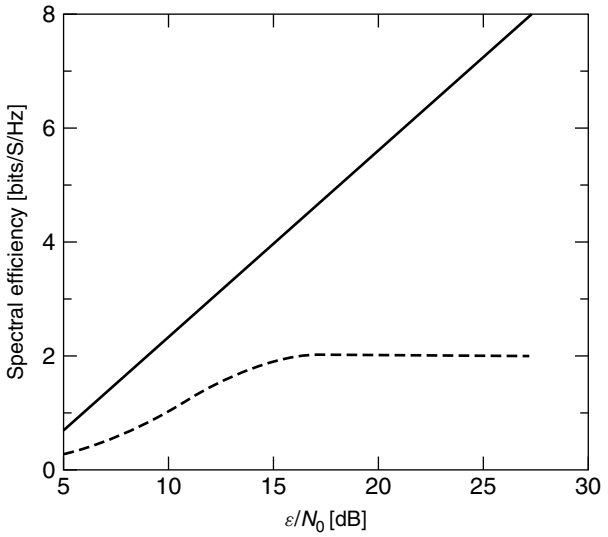


Figure 6. The spectral efficiency of a system using square-law detection. The solid curve corresponds to multilevel transmission constrained only by the average optical power. The dashed curve represents the case of on/off keying. (After Ref. 16.)

the number of transmitted intensity levels. The *spectral efficiency* of an intensity modulated fiberoptic system is defined as the channel capacity per unit of bandwidth. Since a symbol of a temporal duration T occupies a bandwidth equal to at least T^{-1}, the spectral efficiency is given by expression (4) multiplied by the symbol duration T. Figure 6 shows the spectral efficiency extracted from (4) together with the result for on/off keying transmission, as a function of the optical SNR $\mathcal{E}/N_0$. The difference between the two capacities shows the maximum advantage that can be extracted from the use of multilevel intensity modulation. As can be observed in the figure, for typical values of the optical signal to noise ratios (between 10 and 20 dB) multilevel signaling can increase the capacity of an on/off-modulated system by only very small factors. Such small improvement seldom justifies the enormous increase in the complexity and cost that are implied by the use of multilevel optical transmitters and receivers. Notice that the curves shown in Fig. 6 do not take into account the effect of optical nonlinearities, which would undoubtedly reduce the advantage of multilevel modulation even further [17].

4. FIBER TRANSMISSION

The electric field in single-mode optical fibers can be very accurately represented in the form [18]

$$\vec{E}(t,x,y,z) = E(t,z)F(x,y)\exp[-i(\omega_0 t - \beta_0 z)]\hat{u}(t,z) \qquad (5)$$

where z is the propagation axis and x, y are the lateral dimensions of the fiber. The parameter ω_0 is the central frequency of the signal, and β_0 denotes the wavevector at $\omega = \omega_0$. The lateral profile of the beam is $F(x,y)$ and is assumed to remain constant in the process of propagation. The vector $\hat{u}(t,z)$ is a unit polarization vector and lies almost entirely in the x,y plane. In most cases of relevance the transmitted information resides only in the complex envelope of the electric field, which is denoted by the term $E(t,z)$ and is normalized such that $|E(t,z)|^2$ is the optical power in watts. When the effect of polarization-related impairments on system performance is small, it can be shown that the evolution of $E(t,z)$ along the optical fiber is accurately described by the so-called nonlinear Schrödinger equation (NLSE) [18,19]

$$\frac{\partial E}{\partial z} = -i\frac{\beta_2}{2}\frac{\partial^2 E}{\partial t^2} + i\gamma|E|^2E - \frac{\alpha}{2}E \qquad (6)$$

where the first term on the right-hand side describes the effect of chromatic dispersion, the second corresponds to fiber nonlinearity, and the third term corresponds to the scattering losses of the fiber. The derivation of the NLSE relies on the fact that the index of refraction in the fiber can be approximated as $n(\omega,|E|^2) = n(\omega) + n_2|E|^2/A_{\text{eff}}$, where n_2 is the nonlinear refractive index and A_{eff} is the effective cross-sectional area of the beam [18]. Thus the dispersion coefficient is $\beta_2 = \partial^2/\partial\omega^2[\omega n(\omega)/c]$ and the nonlinearity coefficient is $\gamma = \omega_0 n_2/(cA_{\text{eff}})$, where c is the velocity of light in vacuum. The dispersion coefficient can also be expressed in terms of the group velocity v_g in the fiber $\beta_2 = \partial(v_g^{-1})/\partial\omega$ so that it reflects the dependence of the group velocity on the optical frequency. Finally, Eq. (6) is expressed in a delayed timeframe where the average group delay of the propagating pulse is factored out. This means that the time axis t should be interpreted as $t - z/v_g$ in the original representation.

Some insight into the evolution of optical signals in fibers can be obtained by considering cases in which either dispersion or nonlinearity is negligible. First, let us assume that the optical power is low enough to neglect the nonlinear term in the NLSE. Then Eq. (6) can be trivially solved in the Fourier domain yielding

$$\tilde{E}(\omega,z) = \tilde{E}(\omega,0)\exp\left(\frac{i\beta_2\omega^2 z}{2}\right)\exp\left(\frac{-\alpha z}{2}\right) \qquad (7)$$

where $\tilde{E}(\omega,z) = \int_{-\infty}^{\infty} dt\, E(t,z)\exp(i\omega t)$, and ω represents the deviation of the optical frequency from ω_0. Notice that chromatic dispersion does not change the spectral content of the propagating signal since it merely multiplies the original spectrum by a transfer function whose amplitude is equal to 1. Yet it describes a situation where the group velocity varies across the pulse spectrum such that the various frequency components of the pulse tend to separate in time. The most instructive example that illustrates this effect is the case in which the launched pulse is Gaussian: $E(t,0) = A\exp(-t^2/(2\tau_0^2))$. Then the dispersed pulse can be expressed analytically in the time domain $E(t,z) = A(z)\exp[-t^2(1+iC)/(2\tau^2)]$, where C is the chirp parameter $C = \text{sign}(\beta_2)z/L_d$, τ is the width of the dispersed pulse $\tau = \tau_0\sqrt{1+z^2/L_d^2}$, and $L_d = \tau^2/|\beta_2|$ is a parameter with units of length and it describes a characteristic length scale for the effect of dispersion. If we now consider the instantaneous frequency of the pulse, which is equal to minus the time derivative of

its phase $\Delta\omega = -\partial\varphi/\partial t = Ct/\tau^2$, we see that it changes linearly across the pulsewidth. When the fiber dispersion is negative, $\beta_2 < 0$, as is the case in most fibers used for transmission, the leading edge of the pulse consists primarily of the high-frequency content and the trailing edge consists primarily of the low-frequency part. The opposite occurs when $\beta_2 > 0$. Before we conclude the discussion of dispersion as a "standalone" process, we note that it is common practice among optical system engineers to define a different (although equivalent) dispersion coefficient that reflects the dependence of the group velocity on wavelength instead of frequency $D = \partial(v_g^{-1})/\partial\lambda$. It is easy to show that the two coefficients are related to each other by $D = -2\pi c\beta_2/\lambda^2$. It is also common to refer to "normal" and "anomalous" dispersion with regard to the cases $\beta_2 > 0$ ($D < 0$) and $\beta_2 < 0$ ($D > 0$), respectively. Most transmission fibers, as we have noted earlier, are used in the anomalous regime.

To illustrate the effect of fiber nonlinearities we consider the case in which dispersion is negligible. Then the evolution of the electric field assumes the form

$$E(t, z) = E(t, 0)\exp(i\gamma|E(t, 0)|^2 z_{\text{eff}})\exp(-\alpha z/2), \qquad (8)$$

where $z_{\text{eff}} = \int_0^z dz\exp(-\alpha z) = [1 - \exp(-\alpha z)]/\alpha$ is called the *effective length* of fiber, which is the length scale that characterizes the effect of scattering losses. In most relevant cases $\exp(-\alpha L) \ll 1$ so that $z_{\text{eff}} \simeq \alpha^{-1}$. Notice that the effect of the nonlinearity is to modulate the phase of the propagating signal while its intensity remains unperturbed. Therefore this phenomenon is frequently referred to as *self-phase modulation* (SPM). The characteristic length describing the effect of fiber nonlinearities is defined as the effective length of fiber in which the acquired maximum phase shift is equal to 1 radian. It is given by $L_{NL} = (\gamma P_{\text{peak}})^{-1}$, where P_{peak} is the peak power of the optical pulse. The relative significance of dispersion and nonlinearity can therefore be estimated by comparing the characteristic lengths L_{NL} and L_d. As in the case of dispersion, the propagated pulse is characterized by a frequency chirp as its phase becomes time-dependent. The effect on the instantaneous frequency is once again obtained from the derivative of the optical phase $\Delta\omega = -\partial\varphi/\partial t = -\gamma z_{\text{eff}}\partial|E(t, 0)|^2/\partial t$ and it is plotted in Fig. 7 for the case of a typical waveform. As illustrated in the figure, the optical frequency of the leading edge is downshifted and the frequency of the trailing edge is upshifted as a result of the nonlinearity. Consequently, the effect on initially un-chirped pulses will always be that of spectral broadening. Notice, however, that when the launched pulse contains chirp that is opposite to the one caused by nonlinearity (i.e., the leading edge consists of higher frequencies than the trailing edge), the effect of nonlinearity is to equalize the spectrum so that the bandwidth is compressed. This is exactly the situation that occurs when the pulse entering the nonlinear section of fiber is initially predispersed in a linear section of fiber that is characterized by anomalous dispersion.

The general analysis of fiber transmission in the presence of both dispersion and nonlinearity can be handled only numerically. The most efficient and widely used numerical technique for solving the NLSE is the split-step Fourier method [18]. This technique relies on the fact that when the fiber is divided into sufficiently short increments, the propagation of the electric field through each increment can be evaluated in two steps. In the first step the field is propagated through the fiber increment, assuming that it is purely dispersive. In the second step the dispersed signal is propagated through the same increment, assuming that it only contains nonlinearity. The order of the two steps is purely arbitrary and therefore can be reversed without affecting the accuracy of the computation. The representation of the split-step method offers some qualitative insight into the interaction between dispersion and nonlinearity. In particular, we may consider the case of fibers operated in the anomalous regime where the chirp induced by dispersion is opposite in sign to that induced by nonlinearity. Then we may expect that the dispersive step and the nonlinear step in the split-step method will tend to balance each other. A pulse that takes the most advantage of this balance is the optical soliton [19], whose waveform propagates unperturbed along the optical fiber. The electric field envelope of the soliton is given by $E(t, z) = A\,\text{sech}(t/\tau)\exp[-i\beta_2 z/(2\tau)]$, where the pulsewidth τ and the amplitude A are related to each other and to the fiber parameters through $A\tau = \sqrt{|\beta_2|/\gamma}$. It can be shown by direct substitution that the soliton is an exact solution of the NLSE in the absence of scattering losses. In the presence of scattering losses the soliton does not maintain its shape exactly because the strength of the nonlinear effect changes with propagation. Nevertheless, it has been demonstrated that as long as the dispersion length is much greater than the separation between adjacent amplifiers, soliton transmission can exist in a path-averaged sense [20]. Although optical solitons may appear to be a natural solution for fiberoptic transmission, their applicability to optical communications systems

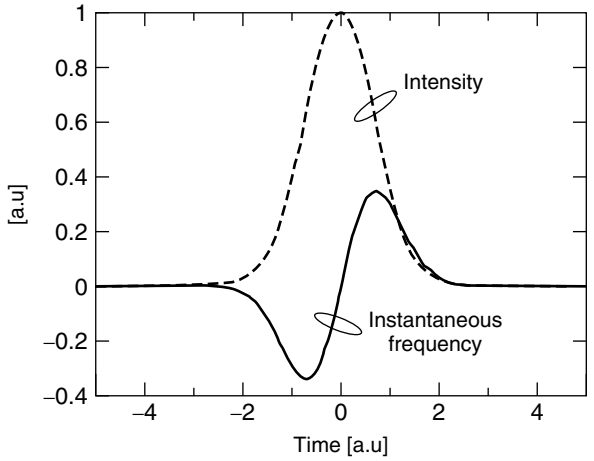


Figure 7. An illustration of SPM in optical fibers. The dashed curve represents the power profile of an optical pulse. The solid curve is the instantaneous frequency shift due to XPM. The shift is proportional to minus the time derivative of the intensity. The leading edge is therefore downshifted in frequency and the trailing edge is upshifted.

remains very limited. What limits their use is primarily the so-called Gordon–Haus effect [21] related to the periodic addition of amplified spontaneous emission noise to the propagating signals, and nonlinear interactions with pulses in adjacent WDM channels, which we review later in this section. Both these effects generate random perturbations of the central frequency of the soliton that are translated into group velocity perturbations due to chromatic dispersion and cause uncertainty in the arrival time of the pulse. Notice that such perturbations of the central frequency occur with all pulses, regardless of their shape. What makes solitons particularly sensitive to frequency perturbations is the fact that they are transmitted without any compensation for chromatic dispersion. Therefore small random variations of the central frequency are translated into enormous variations in the arrival time of the pulse. In spite of very clever ideas for preventing the accumulation of random frequency shifts [22,23], true soliton systems are not used in optical communications. Instead, systems that are designed for reliable long-haul communications are constructed such that most of the dispersion of the link is compensated. Pulses propagating in such systems necessarily change their properties in the process of propagation, and their evolution is strongly dependent on the initial pulseshape and on the amounts of dispersion compensation that are applied along the link. A particularly attractive solution for long-haul transmission is the so-called dispersion-managed soliton, which was originally described in 1995 and 1996 [24,25] and has been extensively studied since then. Dispersion-managed solitons are nearly Gaussian pulses whose evolution is periodic with the period of exactly one span. Specifically, the pulse parameters are chosen such that it undergoes spectral broadening in the first part of its propagation through the fiber span, and then its spectrum is recompressed as it continues to propagate. After applying a properly chosen amount of dispersion compensation at the end of the span, the pulse returns almost precisely to its original waveform. The evolution of dispersion managed solitons can be described quite accurately in a simplified model tracking only a few parameters. These are either the pulse duration and chirp [26], or its bandwidth and frequency dispersion [27]. The reduced models provide very simple numerical methods for extracting the system parameters that are required for dispersion managed soliton transmission. Actual systems parameters are frequently constrained by practical considerations that do not allow matching of the dispersion managed soliton condition exactly. In such cases, when the deviations are small, transmission is often characterized by a periodic "breathing" of the pulse duration and bandwidth with a period of several spans. When the deviation from the dispersion managed soliton requirements is large, the pulses may become completely corrupted, eventually preventing reliable transmission.

In high-channel-count WDM systems one of the most significant impairments to system performance is caused by nonlinear crosstalk between channels [28]. The physical mechanism is quite simple. Every WDM channel is affected by the nonlinear modulation of the refractive index that is caused by the combined intensity of all other channels in the fiber. This process manifests itself in two ways: (1) four-wave mixing (FWM) and the (2) cross-phase modulation (XPM). The *four wave mixing* effect is completely analogous to intermodulation distortions in electronic systems. Any two channels at optical frequencies ω_i and ω_j cause the total optical intensity to oscillate at the frequency difference $\Delta\omega_{i,j} = \omega_j - \omega_i$. These oscillations are imprinted on the refractive index of the fiber such that the propagation of all channels through the fiber is affected. Specifically, a channel at optical frequency ω_k is modulated by the refractive index oscillations and therefore generates new tones at frequencies $\omega_k \pm \Delta\omega_{i,j}$. This interaction occurs between all possible pairs ($\omega_k = \omega_i \neq \omega_j$) and triplets ($\omega_k \neq \omega_i \neq \omega_j$), and therefore, even in systems with a moderate number of channels, an enormous number of new components is created [28]. These components are added as noise to the transmitted channels and may cause significant penalties to system performance. A parameter that strongly affects the efficiency of the FWM process is the chromatic dispersion of the fiber. In the presence of chromatic dispersion, signals at different optical frequencies propagate with different velocities and their phases are acquired at different rates. This implies that the phase with which FWM products are created (which is defined by the phase of the signals that create them) does not match the phase that they acquire during propagation, and therefore their intensity is significantly reduced at the system output. For similar reasons the FWM efficiency also reduces monotonically when the frequency separation between adjacent WDM channels is increased.

The other form of interchannel interference is *cross-phase modulation* (XPM), which is caused by the modulation of the phase of each channel propagating through the fiber by the intensities of all other channels. It is instructive to consider this phenomenon first when the effect of chromatic dispersion is negligible. Then the evolution of the ith channel is expressed analytically as $E_i(t,z) = E_i(t,0)\exp\left[i\gamma z_{\text{eff}}\left(|E_i|^2 + 2\sum_{j\neq i}|E_j|^2\right)\right]$ [18], as can be obtained by direct substitution of the electric field corresponding to multiple channels into Eq. (8) and ignoring the added noise generated by the FWM products. The XPM effect is embodied in the second term in the exponent and similarly to SPM, which we reviewed earlier (represented here by the phase dependence on $|E_i|^2$), XPM affects the instantaneous frequency of the considered channel. Unlike the case of SPM, however, the frequency shift caused by XPM depends on the temporal overlap between the interacting pulses. To clarify this point let us consider the interference between two pulses belonging to two different WDM channels, as illustrated in Fig. 8. Recall that the shift of the optical frequency is given by minus the time derivative of the optical phase. The contribution of XPM to the instantaneous frequency of each pulse is therefore proportional to minus the time derivative of the intensity of the other pulse. As a result, the optical frequency of the leading pulse is reduced by the interaction, whereas the optical frequency of the trailing pulse is increased. In the absence of chromatic dispersion this interference

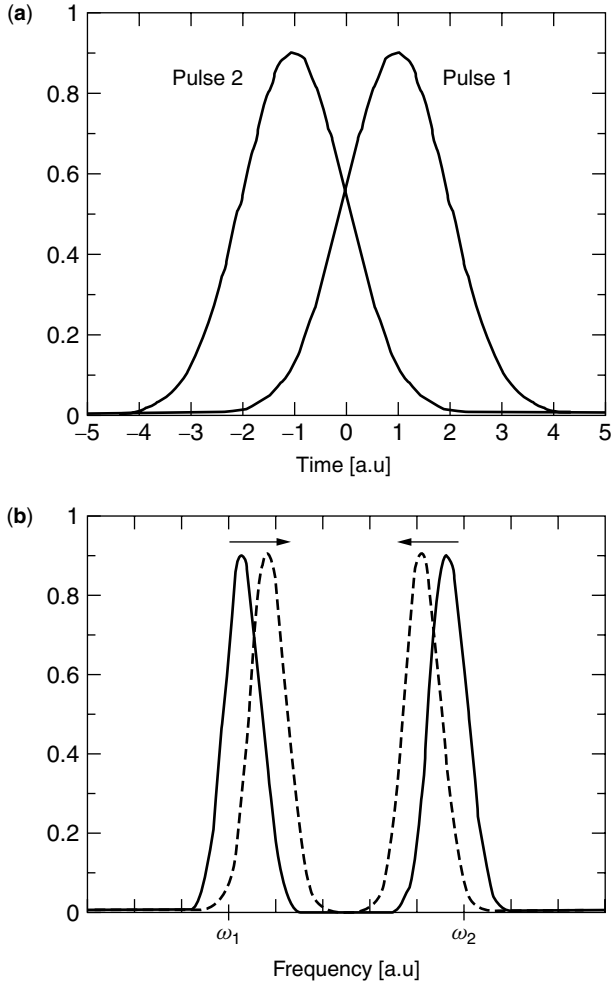


Figure 8. An illustration of a XPM interaction between two pulses transmitted over two different optical frequencies. The solid curves in (**b**) represent the original optical frequencies of the two pulses. The dashed curves represent the spectral shift caused by XPM. The frequency shift experienced by each pulse is proportional to minus the time derivative of the intensity of the other pulse.

turns into a problem only in coherent systems, where information is transmitted on the optical phase, or in intensity-modulated systems using very narrow optical filters that translate the frequency shifts caused by XPM into intensity modulation. Yet zero dispersion fibers are very rarely encountered with WDM transmission since their use would result in significant FWM impairments, as we have discussed earlier. Therefore the more relevant situation to consider is that of transmission in dispersive fibers. The presence of chromatic dispersion modifies the picture presented in Fig. 8 in a number of ways. The most significant modification is that the temporal overlap between the two interacting pulses changes as they propagate along the fiber. In fact, they can overlap only in a fiber section that is no longer than $2\tau_p/(\beta_2\Delta\omega_{i,j})$, where τ_p denotes their duration and $\Delta\omega_{i,j}$ is the difference between their central optical frequencies (so that $\beta_2\Delta\omega_{i,j}$ is the difference between their inverse group velocities). Since the frequency separation $\Delta\omega_{i,j}$ is typically much higher than the spectral width of the individual channels, the effect of dispersion on the individual pulseshapes is negligible within the section of fiber in which the pulses overlap. Finally, the shifts in the optical frequency of the pulses that occur as a result of their interaction are translated into shifts in their group velocities, which lead to uncertainty in their arrival times and cause timing jitter. Notice that chromatic dispersion has two contradicting effects on the significance of XPM-induced penalties. On one hand it is responsible for translating frequency shifts into timing jitter, as we explained earlier, but on the other hand the higher the dispersion, the shorter is the section of fiber in which the pulses overlap, and therefore the magnitude of the frequency shifts becomes smaller [29]. The key difference between these two contradicting mechanisms is that the first is caused by the accumulated dispersion between the section where the pulses interact and the receiver, whereas the latter is related to the local dispersion that characterizes the fiber section in which the nonlinear interaction takes place. Therefore, to reduce the significance of XPM in systems it is beneficial to use high dispersion fiber and compensate for most of the accumulated dispersion at the end of each span.

Other nonlinear phenomena that are not included in the NLSE are the Brillouin and Raman effects, which are both related to the interaction of light with vibrational modes of the medium, or phonons [18]. The Brillouin effect is caused by scattering of light from acoustic phonons and it manifests through the generation of a back scattered wave that carries most of the energy once the incident power exceeds a certain threshold. It is a narrowband phenomenon and can be very effectively avoided by artificially broadening the laser linewidth when necessary [30]. The Raman effect is related to scattering of light from optical phonons. Its most relevant manifestation in communications systems is observed when energy from high-frequency WDM channels is transferred to channels at low frequencies, creating a tilt in the transmitted optical spectrum. The efficiency of the energy transfer induced by Raman scattering peaks for channels separated by approximately 13 THz [31]. This implies that the interacting channels differ significantly in their group velocities as a result of chromatic dispersion, so that the interference between them is averaged with respect to the transmitted data. Consequently, the Raman effect does not constitute a significant mechanism for interchannel crosstalk [28]. The most important application of the Raman effect in fiberoptic systems is in Raman amplifiers, where an intense pump with an appropriately selected optical frequency is launched into the fiber such that it provides gain to the data-carrying signals [32]. The main advantage of this amplification technique is that the gain is distributed along the transmission fiber itself, thereby offering a SNR improvement relative to lumped amplification, consistent with the discussion in Section 2.

5. POLARIZATION-RELATED EFFECTS

Polarization effects in optical systems result from the dependence of the transmission properties of optical fibers and components on the polarization of the transmitted

light. This dependence stems from geometric distortions and mechanical stress that violate the cylindrical symmetry of optical fibers and are created in the process of fabrication and cabling [4]. The most significant consequence of these distortions is *optical birefringence*, which is characterized by the existence of two orthogonal polarization axes $\hat{e}_1$ and $\hat{e}_2$ having different indices of refraction. If we choose to express the electric field as a column vector in the representation of the base defined by the axes of birefringence, the relation between input and output fields is described by

$$\begin{pmatrix} E_1 \\ E_2 \end{pmatrix}_{\text{out}} = \exp(i\beta_0 z) \begin{pmatrix} \exp\left(\frac{i\Delta\beta z}{2}\right) & 0 \\ 0 & \exp\left(\frac{-i\Delta\beta z}{2}\right) \end{pmatrix} \times \begin{pmatrix} E_1 \\ E_2 \end{pmatrix}_{\text{in}} \tag{9}$$

where z is the length of the birefringent section, β_0 is the average wavenumber, and $\Delta\beta$ is the difference between the wavenumbers corresponding to the two birefringent axes. It can be expressed approximately as $\Delta\beta = \Delta n\omega/c$, where Δn is the difference between the indices of refraction of the two axes. The electric field vectors expressed in this way are called *Jones vectors*, and the matrix relating them to each other is called a *Jones matrix*. Notice that the polarization state of a signal is defined jointly by the relative amplitudes of its components and the phase difference between them. Therefore the polarization state of signals that have nonzero components on both $\hat{e}_1$ and $\hat{e}_2$ changes as a result of birefringent propagation. Furthermore, the acquired phase difference and therefore the polarization of the output signal are frequency-dependent, describing a situation that is known as polarization mode dispersion (PMD). If we consider an optical pulse that is launched into a birefringent section of fiber, the output waveform consists of two orthogonally polarized replica of the pulse delayed by $\Delta\beta z$ relative to each other. This situation can be very harmful to the performance of optical communications systems, but it can be easily avoided in a number of ways. For example, one may consider transmitting the signal in a polarization state that coincides with either $\hat{e}_1$ or $\hat{e}_2$ such that no waveform distortions exist at the output. Yet, the more relevant situation to consider is that of a fiber in which the axes of birefringence change in the process of propagation. The most common way of modeling such fibers is by describing them as if they were made of many discrete birefringent sections with random and statistically independent axes of birefringence. This model is justified due to the fact that the correlation length of the birefringence in optical fibers is smaller by several orders of magnitude than the system length [33], so that the details of the local birefringence statistics become irrelevant. In this situation the relation between the input and output states of polarizations is given by

$$\begin{pmatrix} E_1 \\ E_2 \end{pmatrix}_{\text{out}} = \mathbf{T}(i\omega) \begin{pmatrix} E_1 \\ E_2 \end{pmatrix}_{\text{in}} \tag{10}$$

where $\mathbf{T} = \mathbf{T}_N\mathbf{T}_{N-1}\cdots\mathbf{T}_1$, where $\mathbf{T}_i$ is the birefringence matrix of the ith fiber section. Notice that although each individual matrix can be expressed in a diagonal form as the matrix in Eq. (9), they are not diagonal in the same representation and therefore the combined matrix $\mathbf{T}(i\omega)$ is a generic 2×2 unitary matrix that may have an arbitrary dependence on the optical frequency. In this case there is no simple description of the waveform distortions generated by the transmission of a general pulse. An important result can be obtained for the case in which the bandwidth of the launched signal is small relative to the bandwidth that characterizes the frequency dependence of $\mathbf{T}$ [34]. Then Eq. (10) can be approximated in the vicinity of the central frequency of the launched signal ω_0, and it is possible to show that there are two orthogonal states of polarization $\hat{e}_1$ and $\hat{e}_2$ in whose representation equation (10) assumes the form [34]

$$\begin{pmatrix} E_1 \\ E_2 \end{pmatrix}_{\text{out}} \simeq e^{i(\omega-\omega_0)\tau_0/2} \begin{pmatrix} e^{-i(\omega-\omega_0)\tau/2} & 0 \\ 0 & e^{i(\omega-\omega_0)\tau/2} \end{pmatrix} \times \mathbf{T}(i\omega_0) \begin{pmatrix} E_1 \\ E_2 \end{pmatrix}_{\text{in}} \tag{11}$$

where $\mathbf{T}(i\omega_0)$ is the polarization rotation corresponding to the central frequency and τ_0 is the average, polarization-independent time delay. Notice that apart from the frequency-independent operator $\mathbf{T}(i\omega_0)$, Eq. (11) is identical to Eq. (9), which represents the case of birefringence with fixed axes. Therefore a pulse with nonzero components on $\hat{e}_1$ and $\hat{e}_2$ comes out of the fiber consisting of two orthogonally polarized replica of itself delayed by τ relative to each other. The time delay τ is known as the *differential group delay* (DGD), and the polarization states $\hat{e}_1$ and $\hat{e}_2$ are known as the *principal states of polarization* (PSPs). Equation (11) represents what is known as the first-order PMD approximation because it relies on the first-order expansion of the matrix $\mathbf{T}$ with respect to frequency. This description has been shown to be very useful in predicting the kind of penalties induced by PMD in fiberoptic systems in a reasonably broad range of parameters [4,35]. The further analysis of PMD requires tools whose description is beyond the scope of this article [36]. We will therefore limit ourselves to stating some of the most relevant results concerning this phenomenon. One obvious observation is that the DGD and the PSPs are stochastic processes with respect to both the central frequency and the position along the fiber. It can be shown that the differential group delay at any fixed frequency is a Maxwell distributed parameter and its mean value $\langle\tau\rangle$ is proportional to the square root of the fiber length [37]. The bandwidth that characterizes the frequency dependence of the transfer matrix $\mathbf{T}(i\omega)$ can be estimated by deriving frequency autocorrelation functions of the DGD and the principal states [38–40], which indicate that the bandwidth of all phenomena related to PMD is of the order of $\langle\tau\rangle^{-1}$. The nature of higher-order distortions that are caused by PMD is still one of the most active topics of research in optical communications. With the rapid increase in data rates transmitted over modern systems, the importance of this topic is growing accordingly. One important general property of PMD that is related to the rapidly increasing data rates is that its

effect can be rigorously scaled on the basis of the mean DGD of the fiber [40]. Thus for example a 10-Gbps channel transmitted over a system with 10 ps mean DGD experiences the same waveform distortions as a 40-Gbps channel transmitted over a system with a mean DGD of 2.5 ps.

In addition to the effect of birefringence, which limits the performance of optical systems by distorting the transmitted waveforms, optical systems can also be penalized by the existence of polarization-dependent loss (PDL) along the optical link. As is suggested by its name, PDL describes a situation in which there exist two orthogonal polarization components characterized by different attenuation. Whereas PMD is caused primarily by the birefringence of optical fibers, the main source of PDL is the imperfection of inline optical components such as isolators, dispersion compensating devices, and optical switches. In principle, PDL may contribute to the generation of waveform distortions through a nontrivial interaction with the fiber birefringence [41]. In practice, however, considering typical system parameters, the effect of PDL on the waveform is usually small. Instead, it effects the performance of optical communications systems by penalizing the optical signal-to-noise ratio of the received signals [42,43]. The deterioration of the signal-to-noise ratio occurs in two ways. The first is caused by the fact that components having PDL may attenuate the propagating signal by more than the average attenuation and therefore the SNR is reduced accordingly [42]. The second results from the fact that in the presence of PDL, noise that is originally emitted into a state of polarization that is orthogonal to the signal is coupled into the signal's polarization [43] and therefore mixes with the signal at photodetection. The significance of PDL is limited primarily to long-haul terrestrial systems using affordable components.

The preceding description of polarization-related phenomena assumed that they could be treated separately from nonlinear propagation. This assumption is characteristic of the vast majority of studies that have been reported in the literature so far. It is reasonable in systems operating over high PMD links [4], where PMD is the most significant limiting factor. A more general treatment of polarization related effects is based on the so-called coupled nonlinear Schrödinger equations [44–46], which can simultaneously take into account the effects of birefringence, PDL, dispersion, and nonlinear propagation. Such simultaneous treatment of all propagation phenomena is particularly important in long-haul systems where proper operation requires fine control of the dispersion map. In those cases even moderate levels of PMD can disturb the system and cause significant penalties to performance.

Acknowledgment

The author is pleased to acknowledge A. Mecozzi and J. P. Gordon for the multiple discussions that were conducted on the topics included in this article.

BIOGRAPHY

Mark Shtaif received his M.Sc. and Ph.D. degrees in electrical engineering at the Technion in 1993 and 1997, respectively. In 1997 he joined the Light-wave Networks Research Department at AT&T Labs Research as a Senior and then Principal Member of Technical Stuff. In AT&T his work was centered around the modeling and characterization of optical fiber communication systems, focusing on propagation effects in optical fibers, including fiber nonlinearities, polarization mode dispersion, special modulation formats, and interaction of signals and noise. During his employment in AT&T he served as a technical consultant to the AT&T business units, on the evaluation of fiberoptic technologies. In December 2001 he became a Principal Architect in Celion Networks, an optical networking company, where he worked on the analysis and design of long-haul optical transmission systems. In April 2002 Dr. Shtaif joined the faculty of the Engineering Department in Tel-Aviv University, where he conducts research and teaches courses in the area of optical communications.

BIBLIOGRAPHY

1. R. J. Mears, L. Reekie, I. M. Jauncey, and D. N. Payne, Low noise erbium doped fibre amplifier operating at 1.54 μm, *Electron. Lett.* **23**: 1026 (1987).
2. E. Desurvire, J. R. Simpson, and P. C. Becker, High-gain erbium doped fiber amplifier, *Opt. Lett.* **12**: 888 (1987).
3. A. H. Gnauck and R. M. Jopson, Dispersion compensation for optical fiber systems, in I. P. Kaminow and T. L. Koch, eds., *Optical Fiber Telecommunications IIIA*, Academic Press, San Diego, CA, 1997, Chap. 7.
4. C. D Pool and J. A. Nagel, Polarization effects in lightwave systems, in I. P. Kaminow and T. L. Koch, eds., *Optical Fiber Telecommunications IIIA*, Academic Press, San Diego, CA, 1997, Chap. 6.
5. K. Shimoda, H. Takahasi, and C. H. Townes, Fluctuations in the amplification of quanta with application to Maser amplifiers, *J. Phys. Soc. Jpn.* **12**: 686 (1957).
6. H. A. Haus and J. A. Mullen, Quantum noise in linear amplifiers, *Phys. Rev.* **128**: 2407 (1962).
7. H. Kogelnik and A. Yariv, Considerations of noise and schemes for its reduction in laser amplifiers, *Proc. IEEE* **52**: 165 (1964).
8. E. Desurvire, *Erbium Doped Fiber Amplifiers: Principles and Applications*, Wiley, New York, 1994, Chap. 2.
9. H. A. Haus, Noise figure definition valid from RF to optical frequencies, *IEEE J. Select. Top. Quant. Electron.* **6**: 240 (2000).
10. J. P. Gordon and L. F. Mollenauer, Effects of fiber nonlinearities and amplifier spacing on ultra-long distance transmission, *J. Lightwave Technol.* **9**: 170 (1991).
11. L. Kazovsky, S. Benedetto, and A. Willner, *Optical Fiber Communications Systems*, Artech House, Norwood, MA, 1996, Chap. 4.
12. G. P. Gordon and L. F. Mollenauer, Phase noise in photonic communications systems using optical amplifiers, *Opt. Lett.* **15**: 1351 (1990).
13. P. A. Humblet and M. Azizoglu, On the bit error rate of lightwave systems with optical amplifiers, *J. Lightwave Technol.* **9**: 1576 (1991).

14. J. G. Proakis, *Digital Communications*, McGraw-Hill, New York, 2001.

15. C. E. Shannon, A mathematical theory of communication, *Bell. Syst. Tech. J.* **27**: 379 (July 1948); 623 (Oct. 1948).

16. A. Mecozzi and M. Shtaif, On the capacity of intensity modulated systems using optical amplifiers, *IEEE Photon. Technol. Lett.* **13**: 1029 (2001).

17. P. P. Mitra and J. B. Stark, Nonlinear limits to the information capacity of optical fibre communications, *Nature* **411**: 1027 (2001).

18. G. P. Agrawal, *Nonlinear Fiber Optics*, Academic Press, San Diego, CA, 1989.

19. L. F. Mollenauer, J. P. Gordon, and P. V. Mamyshev, Solitons in high bit-rate long-distance transmission, in I. P. Kaminow and T. L. Koch, eds., *Optical Fiber Telecommunications IIIA*, Academic Press, San Diego, CA, 1997, Chap. 12.

20. L. F. Mollenauer, J. P. Gordon, and M. N Islam, Soliton propagation in long fibers with periodically compensated loss, *IEEE J. Quant. Electron.* **22**: 157 (1986).

21. J. P. Gordon and H. A. Haus, Random walk of coherently amplified solitons in optical fiber transmission, *Opt. Lett.* **11**: 665 (1986).

22. A. Mecozzi, J. D. Moores, H. A. Haus, and Y. Lai, Soliton transmission control, *Opt. Lett.* **16**: 1841 (1991).

23. L. F. Mollenauer, J. P. Gordon, and S. G. Evangelides, The sliding frequency guiding filter, an improved form of soliton jitter control, *Opt. Lett.* **17**: 1575 (1992).

24. M. Suzuki et al., Reduction of Gordon-Haus timing jitter by periodic dispersion compensation in soliton transmission, *Electron. Lett.* **31**: 2027 (1995).

25. N. J. Smith et al., *Electron.Lett.* **32**: 54 (1996).

26. J. N. Kutz, P. Holmes, S. G. Evangelides, and J. P. Gordon, Hamiltonian dynamics of dispersion managed breathers, *J. Opt. Soc. Am.* **15**: 87 (1998).

27. J. P. Gordon and L. F. Mollenauer, Scheme for the characterization of dispersion managed solitons, *Opt. Lett.* **24**: 223 (1999).

28. F. Forghieri, R. W. Tkach, and A. R. Chraplyvy, Fiber nonlinearities and their impact on transmission systems, in I. P. Kaminow and T. L. Koch, eds., *Optical Fiber Telecommunications IIIA*, Academic Press, San Diego, CA, 1997, Chap. 8.

29. M. Shtaif, An analytical description of cross phase modulation in dispersive optical fibers, *Opt. Lett.* **23**: 1191 (1998).

30. D. A. Fishman and J. A. Nagel, Degradations due to stimulated Brillouin scattering in multigigabit intensity-modulated fiber-optic systems, *J. Lightwave Technol.* **11**: 1721 (1993).

31. R. H. Stolen and E. P. Ippen, Raman gain in glass optical waveguides, *Appl. Phys. Lett.* **22**: 294 (1973).

32. P. B. Hansen et al., Capacity upgrades of transmission systems by Raman amplification, *IEEE Photon. Technol. Lett.* **9**: 262 (1997).

33. A. Galtarossa, L. Palmieri, M. Schiano, and T. Tambosso, Measurement of birefringence correlation length in long, single-mode fibers, *Opt. Lett.* **26**: 962 (2001).

34. C. D. Poole and R. E. Wagner, Phenomenological approach to polarization dispersion in long single mode fibers, *Electron. Lett.* **22**: 1029 (1986).

35. H. Kogelnik, R. M. Jopson, and L. E. Nelson, Polarization mode dispersion, in I. P. Kaminow and T. Li, eds., *Optical Fiber Telecommunications Ivb: Systems and Impairments*, Academic Press, San Diego, 2002, Chap. 15.

36. J. P. Gordon and H. Kogelnik, PMD fundamentals, *Proc. Nat. Acad. Sci.* **97**: 4541 (2000).

37. G. J. Foschini and C. D. Poole, Statistical theory of polarization dispersion in single mode fibers, *J. Lightwave Technol.* **9**: 1439 (1991).

38. M. Karlsson and J. Brentel, Autocorrelation function of the polarization mode dispersion vector, *Opt. Lett.* **24**: 939 (1999).

39. M. Shtaif, A. Mecozzi, and J. A. Nagel, Mean square magnitude of all orders of polarization mode dispersion and the relation with the bandwidth of the principal states, *IEEE Photon. Technol. Lett.* **12**: 53 (2000).

40. M. Shtaif and A. Mecozzi, Study of the frequency autocorrelation of the differential group delay in fibers with polarization mode dispersion, *Opt. Lett.* **25**: 707 (2000).

41. B. Huttner, C. Geiser, and N. Gisin, Polarization-induced distortions in optical fiber networks with polarization mode dispersion, *IEEE J. Select. Top. Quant. Electron.* **6**: 317 (2000).

42. E. Lichtman, Limitations imposed by polarization dependent gain and loss on all-optical ultra-long communications systems, *J. Lightwave Technol.* **13**: 906–913 (1995).

43. M. Shtaif, A. Mecozzi, and R. W. Tkach, Noise enhancement caused by polarization dependent loss and the effect of gain equalizers, *Proc. Optical Fiber Communications Conf.*, Anaheim, CA, 2002, Paper TuL1.

44. C. R. Menyuk, Nonlinear pulse propagation in birefringent optical fibers, *IEEE J. Quant. Electron.* **23**: 174 (1987).

45. S. G. Evangelides, L. F. Mollenauer, J. P. Gordon, and N. S. Bergano, Polarization multiplexing with solitons, *J. Lightwave Technol.* **10**: 28 (1992).

46. P. K. Wai, W. L. Kath, C. R. Menyuk, and J. W. Zhang, Nonlinear polarization mode dispersion in optical fibers with randomly varying birefringence, *J. Opt. Soc. Am. B* **14**: 2967–2979 (1997).

MODEMS

RAVI BHAGAVATHULA
HYUCK KWON
Wichita State University
Wichita, Kansas

1. INTRODUCTION

Transmission of data between two devices requires the usage of a transmitter, a receiver, and a transmitting media that provides a path between the transmitter and the receiver. According to the manner in which data are transmitted, there are two fundamental modes of transmission: (1) parallel and (2) serial.

In a parallel mode of transmission, data are transmitted one byte (or character) at a time. This mode of transmission requires a minimum of 8 lines, with additional lines for control signaling, for transmitting the 8 bits (or one byte) of data from the transmitter to the receiver. This

transmission method yields a very high data rate at the expense of increased costs due to the presence of a large number of cables between the communicating devices. Hence, it is used in communication between computers and peripheral units where cable distances are relatively short and data transfers must occur rapidly (such as between a printer and a computer).

The parallel mode of transmission becomes increasingly expensive as the distance between the two communicating devices increases relative to the increase in the cost of the cables. An alternative to parallel transmission is the serial mode of transmission, wherein the data are transmitted in sequence over one line, that is, one bit at a time. Instead of requiring additional lines for control signals, a preset sequence of bits can be used for a similar purpose, allowing a two-wire circuit with one wire serving as an electrical ground to be used for data transmission. Since the public switched telephone network (PSTN) already provides a two-wire facility for voice transmission, it is quite logical for serial transmission to utilize the available infrastructure for a cost-effective transmission mechanism.

To communicate over serial lines, the terminal devices need to convert the parallel data into a serial datastream. A *universal asynchronous receiver/transmitter* (UART) on the terminal device usually handles this, and the resulting serial datastream is transmitted/received using a common serial interface. However, there is a basic incompatibility between the digital signals transmitted by a terminal device and the analog signals transmitted by a PSTN line since the PSTN was originally designed to carry only voice signals. Although digital signals can be transmitted over an analog telephone line, the digital pulse-distorting effects of resistance, inductance, and capacitance on the analog PSTN line limit their transmission distance. Further, the presence of analog amplifiers to boost analog voice signal levels in the PSTN pose additional problems with digital data since the analog amplifier would boost the digital signal along with the distortions and would, therefore, increase the distortion in the digital data transmission.

Because of the incompatibilities between the digital signals produced by terminal devices and the analog signals that telephone lines were designed to carry, a conversion device is required to enable digital signals to be carried on an analog transmission medium. This conversion device is a modem, a contraction of the term modulator–demodulator. The modulator portion of the device converts (or modulates) the digital signals into analog signals for transmission over the PSTN line, while the demodulator portion of the device converts (or demodulates) the analog signal into digital format. Therefore, the modulator portion of the modem can be considered to be the transmission component of a communication system, and the demodulator can be considered to be the receiver component of a communication system.

With the emergence of digital telephony, some portions of the PSTN are designed to carry voice signals in a digital format. Usage of services such as the Integrated Services Digital Network (ISDN) allows PSTN subscribers for end-to-end voice and data communication in digital format. These digital networks use a bipolar signaling scheme for transmission over twisted-pair cable. A digital modem is therefore utilized to convert the unipolar signals generated by the terminal devices to a bipolar format used by the digitized PSTN.

Modems, depending on the type of datastreams they operate on, work in either an asynchronous or a synchronous mode. In an asynchronous mode of operation, often referred to as a *start/stop* transmission, each character is encoded into a series of pulses. The transmission is started by a start pulse followed by the encoded character (a series of pulses). The receiver is notified of the completion of the transmission of a character by the transmission of a stop pulse that may be equal to or longer than the start pulse depending on the transmission code being used.

In a synchronous mode of operation, a group of characters are transmitted in a continuous bitstream. Modems located at each end of the transmission medium normally provide a timing signal or clock to establish the data transmission rate and hence enable the devices attached to the modems to identify the appropriate characters as they are being transmitted or received. Before the data transmission is initiated, the transmitting and the receiving devices must establish synchronization between themselves. To keep the receiving clock in step with the transmitting clock for the duration of a bitstream representing a large number of consecutive characters, the data transmission is preceded by the transmission of a special set of synchronization characters. An error-free data transmission after the synchronization process is achieved by using an error detection scheme known as *cyclic redundancy check* (CRC). This mode of serial data transfer yields a much higher data rate at the expense of complex circuitry because the receiver must remain in phase with the transmitter for the duration of the transmitted group of characters.

2. MODEM OPERATION

A modem connects a PC (or a computing device) with the outside world through a series of cables. The connection between the modem and the PC is usually accomplished by a serial cable (in case of an external modem) or through the system bus itself (in case of an internal modem). The modem speaks with the outside world through a telephone line.

In the common scenario of a PC speaking with the outside world through a modem, the PC is referred to as a DTE (data terminal equipment) while the modem is referred to as a DCE (data communication equipment). An exception is the case of an internal modem wherein the concept of a DTE does not exist since the definitions of DTE and DCE are plausible with respect to a RS-232 connection (and an internal modem does not employ any RS-232-type connections since it is a bus-connected device and not a serial-interface-connected device).

The speeds at which a DTE can transmit information to a modem are usually much larger than the speeds at which

the modem can transmit that information to the outside world. This speed difference warrants the existence of a number of mechanisms to ensure the timely and accurate exchange of information.

In an effort to bridge the communication speed differences, various data compression schemes are employed. The data compression schemes currently in use are the MNP-5 and the V.42bis. MNP-5 is derived from Microcom Network Protocol standard (devised by Microcom, now acquired by Compaq) and yields a data compression rate of up to 2 : 1. However, MNP-5 accumulates a large amount of overhead while compressing data. Further, it is rather inefficient in transmitting precompressed files since it cannot sense the need for compression. Using this scheme with a precompressed file usually results in the transmission of a larger file as compared to the original compressed file.

V.42bis is a data compression standard approved by ITU-T (International Telecommunication Union — Telecommunication Sector) [1], which yields a compression ratio of up to 4 : 1. It builds on the advantages of the MNP-5 protocol and includes the ability to sense when compression is required. This makes it more efficient when transmitting precompressed files.

It should be noted that the two standards, MNP-5 and V.42bis, are both exclusive in nature; that is, they cannot be used at the same time. Further, for optimal performance, the DTE-DCE communication link (usually the terminal and the modem connection) should be able to sustain the data rate that these compression schemes afford. In the case of the MNP-5, since the compression rate is 2 : 1, for every bit that the modem sends out, the DTE is required to transmit 2 bits' worth of information. Therefore, the bit rate of the DTE should be 2 times that of the DCE. In the case of V.42bis, the DTE speed should be 4 times that of the DCE. For example, if your modem is operating at 14,400 bps and no data compression schemes are being used, the DTE speed would need to be set at 14,400 bps (bits per second). However, if MNP-5 were to be used, the DTE speed would need to be set at 28,000 bps (2 × DCE speed) since the MNP-5 supports a compression ratio of 2 : 1. In the case of V.42bis, the DTE speed would need to be 57,600 bps (4 ∗ DCE speed) as V.42bis supports a compression ratio of 4 : 1.

A convenient way in which these speed differences are quoted in modem terminology is through the concept of a baud. A *baud* is defined as the rate at which information is transmitted. In contrast, *bit rate* is defined as the rate at which bits are transmitted. Put another way, a baud could be defined as one pulse (or signal) interval in a carrier signal while bit rate is the number of bits that are transmitted per second (or signal). The relation between baud and bit rate is given below:

$$\text{Bit rate (in bits per second)} = \text{baud} \times \text{bits per baud}$$

With increasing transmitting speeds, the emphasis on better error control mechanisms has been increasing for the accurate exchange of information. Two distinct error control schemes that are used in most modern modems are MNP-4 and V.42. MNP-4 includes the functionalities of MNP-2 and MNP-3 while V.42 employs LAP-M (Link Access Protocol — Modem) protocol as the primary error control mechanism and reverts to MNP-4 as a backup. These error control schemes retransmit corrupted data using 16–32-bit CRCs. V.42 is an ITU-T specification that yields slightly better performance than MNP-4. In V.42, the primary error control mechanism is LAP-M. Data are grouped together in terms of frames, and these frames are transmitted over the communication channel along with a CRC header for error control. In case of LAP-M, each frame has a size of 128 bytes, and up to 15 frames (by default) can be sent without waiting for an acknowledgment from the receiver. This translates to a storage requirement of $128 \times 15 = 1920$ bytes at the transmitting end for accomplishing error-free transmission since the transmitter would need to retransmit all 15 frames if the transmitter receives a negative acknowledgment from the receiver. V.42 employs 16- or 32-bit CRC fields (although 32-bit CRC is more common these days.)

Because of the presence of many operating speeds, it is quite possible that the operating speed of one modem may be more than that of another modem involved in a typical end-to-end connection between two communicating terminals. In such cases, the receiving modem needs to be able to inform the transmitting modem to pause before it can fully process the data it received. This is accomplished using flow control mechanisms. Flow control mechanisms can be broadly classified as either software-based or hardware-based.

In *software-based* flow control mechanisms (also referred to as XON/XOFF mechanisms), the receiver modem sends a special signal (usually a Control-S character) to the transmitting end requesting the transmitter modem to pause for a while. The transmitter modem stops sending any new information to the receiver modem until it receives another special signal from the receiver modem (which is usually a Control-Q character) informing the transmitter modem that it can resume sending data. The advantage of this scheme is that no additional hardware support is required since the pause/resume signals are handled by the communication software (or the firmware in the modem). The disadvantage of this scheme is that the presence of noise in the transmission media can result in the loss of the pause/resume signals and therefore affect the operation of the modems. If the pause signal were lost, the transmitter modem would keep transmitting data at a rate that the receiver modem cannot handle, resulting in overruns at the receiving modem. If the resume signal is lost, the transmitter modem would never know when to resume transmission of data and therefore the transmission media would remain silent forever. Because of the transmission of special characters to signify pause/resume actions during transmission, XON/XOFF flow control mechanism should not be used for the transfer of binary data since the modems could falsely interpret the presence of Control-S character in the original binary file as a signal to pause data transmission.

In sharp contrast to software-based flow control schemes, *hardware-based* flow control schemes depend on special hardware support for ensuring proper control over

the flow of data. This is also referred to as RTS/CTS (ready to send/clear to send) flow control mechanism. In the case of external modems, a specific wire in the serial cable (that is used to connect the terminal to the modem) is used for exchanging flow control information. In internal modems, in the absence of a serial cable, flow control functionality is built into the modem itself.

Data transfer (more commonly known as *file transfer*) protocols exist for the transmission of binary data between two modems over a communication channel (which is usually a telephone line). Commonly used data transfer protocols include Xmodem, Xmodem-CRC, Xmodem-1K, Ymodem, and Zmodem. In the *Xmodem* transfer protocol, binary data are transmitted in 128-byte chunks and a checksum is appended to these 128-byte blocks so that the receiver can verify the integrity of the received block and intimate the transmitter of the status of its reception. If the receiver determines any errors in the reception, the transmitter modem would resend the entire 128-byte block.

Xmodem-CRC adds CRC functionality to the basic Xmodem protocol, while Xmodem-1K transfers data in terms of 1-kbyte data chunks as against the standard 128-byte data chunks used by the standard Xmodem protocol. The *Ymodem* protocol is quite similar to the *Xmodem-1K* protocol and is seldom used on noisy communication channels (like telephone lines) due to its inability to perform efficiently in such environments. Ymodem-G, an improvement on the original Ymodem protocol, yields slightly faster data transfer rates by eliminating software error control mechanisms and relying on the underlying hardware to perform the required error control operations.

Zmodem is the most widely used data transfer protocol over dialup connections because of its improved resilience to noisy environments and the higher data rates. It employs 32-bit CRC fields and does not wait for an acknowledgment from the receiver before it transmits the next block of data.

3. MODULATION TECHNIQUES AND MODEM STANDARDS

Because of the comparatively smaller bandwidths that are offered by present-day dialup connections, modems employ a modulation scheme to transmit more information at the same bit rate. This is accomplished by converting data into *symbols* and transmitting the symbols over the communication channel. The conversion of a datastream into a symbol stream is carried out by using an appropriate modulation scheme. The usage of these modulation schemes leads to a much better utilization of bandwidth because more information (in terms of data bits) can be inserted into specific *symbols* that are transmitted over the communication channel.

The most common modulation scheme is AM (amplitude modulation) [2]. This forms the basis for a few more advanced modulation schemes like QAM (quadrature amplitude modulation) [2]. In AM, symbols are defined in terms of the amplitude of the original signal and these symbols are transmitted over a carrier through the analog communication channel. A major disadvantage of AM is the fact that as the amplitude of the signal decreases, it becomes increasingly difficult to separate the signal from noise in the communication channel.

QAM is an improvement over AM in which information is encoded on the basis of the deviations in the phase and amplitude of the carrier wave. This so-called *two-dimensional* encoding leads to a greater encoding efficiency and, therefore, a higher data rate.

TCM (trellis-coded modulation) [3] is based on the QAM scheme. In TCM, additional bits are added to each *symbol* to accomplish *forward correction*. This leads to a better error control ability and bit errors introduced into the communication process can be effectively reduced.

FM (frequency modulation) [2] is the frequency counterpart for the AM scheme wherein the modulation is accomplished in terms of the frequency rather than the amplitude. While this modulation scheme is used more widely for radio broadcasts, its application in dialup connections is not widespread. A variation of FM modulation is FSK (frequency shift keying), which was designed primarily for transmission of data across a telephone line. In this scheme, the presence of bit "1" is represented by a specific frequency tone and the presence of bit "0" is represented by another specific frequency tone. To afford two-way communication, FSK allows the specification of two different sets of frequency tones.

PSK (phase shift keying) is similar in principle to FSK, with the sole difference that variations in phase of a constant frequency carrier are used to determine the bit values in the original datastream. A signal with unchanged phase is used to signify the presence of a bit whose value is the same as that of the previous bit. A 50% change in the phase is used to signify a bit value that is different from the previous bit. Differential PSK (DPSK) is a refinement of PSK wherein the changes in phase are determined by comparing the phase of the current state with that of its previous state.

PCM (pulse code modulation) is a modulation scheme wherein analog data are encoded into a specific number of bits (usually 8 bits) for transmission over a communication channel. This is, strictly speaking, not a true modulation scheme since a carrier is not employed at all. The encoding of analog information into digital format is accomplished using a quantizer and a sample-and-hold circuit.

Several modem standards have been introduced that allow modems to exchange data universally. While it is not possible to discuss every modem standard in this document, a few representative standards are discussed here.

Bell 103 is one of the older standards that allow data to be transmitted and received at a rate of 300 bps (bits per second). It employs FSK modulation and uses 1 bit to represent a baud (i.e., bit rate is equal to baud rate). Bell 202 is considered an improvement over the Bell 103 as it supports a 1200 bps data rate using the same FSK modulation technique as employed by Bell 103.

With the widespread usage of modems across the globe, the need arose for a set of modem standards that could facilitate modems throughout the world to communicate with each other. CCITT (later known as *ITU-T*) undertook to formulate a set of universally applicable standards to

this effect. One of the earlier CCITT standards for data communication was the CCITT V.21 [4], which allowed a data rate of 300 bps using FSK modulation (quite similar in operation to the Bell 103 standard). CCITT V.22 used DPSK to obtain data rates of 1200 bps at a baud rate of 600 baud (i.e., the number of bits per baud was set to be equal to 2). This standard is similar to the Bell 212A.

More advanced standards were later released by CCITT (viz., V.32 and V.32bis [5]) that allowed data rates of up to 14,400 bps using different modulation techniques such as TCM and QAM. ITU-T V.34 [6] is a more recently established standard that allows modems to transfer data at rates up to 33,600 bps.

V.34 is currently the fastest end-to-end analog modem standard. Because of the dependence of the more advanced standards such as V.90 on the V.34 standard, let us take a closer look at the details of the V.34 analog telephony standard.

Modems supporting V.34 can sustain data transmission capacities of 2400–28,800 bps. A feature referred to as *line probing* was introduced in this standard to allow modems to identify the capacities and quality of the phone landline and adjust themselves to allow, for each individual connection, the most optimal data transmission rate. V.34 also supports a synchronous auxiliary channel with a data rate of 200 bps that could be used in tandem with the primary data channel for signaling information.

The bandwidth offered by a phone line is around 3–4 kHz, and the maximum symbol rate that is supported by V.34 is 3429 symbols per second. The operation of V.34 near the theoretical limits of the phone line spurred the design engineers to incorporate a mechanism within V.34 to autonegotiate the available bandwidth on a phone line and adjust the data transmission rates accordingly. A new handshake protocol, called as V.8 mode negotiation handshake, was introduced to enable two V.34 compatible modems to exchange feature and mode negotiation information via V.21 standards (300 bps FSK modulated communication). V.8 mode is used by the two V.34 modems to identify them with other telephone network equipment. It is also used to determine whether the call is destined for a data or facsimile operation. Negotiation of the available modes of modulation is also accomplished along with ability to support V.42 and V.42bis standards [1]. During this handshake, the modems send a series of tones to each other, at specific frequencies and known signal levels. The received signal level is employed in the computation of the maximum possible available bandwidth for communication.

Line probing is employed immediately after V.8 handshake to determine parameters such as the optimal bandwidth and carrier frequency, preemphasis filters, and optimal output power level to be used during communication. Table 1 lists the symbol rates, carrier frequencies, and supported data rates as defined in V.34 that are implemented in the 3Com OfficeConnect series of modems. For every symbol rate [except 3429 symbols per second (sps)], the modem can select one of the two available carrier frequencies.

A preemphasis filter is usually employed in a modem to remove amplitude distortions that could creep into a phone line, by suitably shaping the transmitted signal's spectrum. V.34 supports 10 preemphasis filters, and the appropriate filter is selected during the line probe operation.

Table 1. V.34-Supported Symbol Rates, Carrier Frequencies, and Bandwidths (Compatible with 3Com OfficeConnect Series of Modems)

Symbol Rate	Minimum Bit Rate	Maximum Bit Rate	Carrier Frequency	Required Bandwidth
2400	2400	21,600	1600	400–2800
			1800	600–3000
2743	4800	24,000	1646	274–3018
			1829	457–3200
2800	4800	24,000	1680	280–3080
			1867	467–3267
3000	4800	26,400	1800	300–3300
			2000	500–3500
3200	4800	28,800	1829	229–3429
			1920	320–3520
3429	4800	28,800	1959	244–3674

Compared to the 2-D TCM schemes employed by V.32bis, V.34 lets the modems select any one of the three available 4-D TCM schemes. This allows for a more robust error-correction scheme for accurate data transmission.

When two modems lose synchronization with each other, a feature called a retrain is activated. During retraining, the modems do not send any data across the network since they suspend all operations and re-negotiate the connection once again. Unlike the earlier standards, retraining in V.34 is accomplished using the receiver modem's timer. This leads to a reduced (and variable) retrain time as compared to a large, fixed retrain time for the earlier standards.

With the emergence of faster terminals, the need arose for standards that would enable PCs to communicate more rapidly than the allowed 33,600 bps (as with ITU-T V.34). Two contemporary standards evolved to meet these requirements: the Rockwell/Lucent K56 standard and the USR X2 standard.

Traditionally, communication between modems is carried out on the assumption that both ends of a modem conversation have an analog connection to the telephone network. Therefore, data from a terminal are converted from digital format to analog format and transmitted over the PSTN (at speeds of up to 33,600 bps as supported by V.34). At the receiving end, a modem translates the analog signals into digital data that the receiving terminal can process.

In contrast, the Rockwell/Lucent K56 and USR X2 protocols assume that one end of the modem conversation has a digital connection to the telephone network (usually through a digital modem like ISDN). Since Internet service providers (ISPs) usually transmit data in digital format across networks, the digital end of the modem conversation is typically that of an ISP.

Since one end of the modem conversation is digital in nature, the downstream traffic (traffic from the ISP to a user's modem) is digitally modulated using PCM yielding data rates of up to 56,000 bps. Upstream traffic (traffic

from a user's modem to the ISP) still proceeds at a rate of 33,600 bps (using V.34 standards). Therefore, the K56 and the X2 standards are asymmetric in nature since they offer different data rates for different directions of data transfer.

Because of the incompatibility of the K56 and the X2 standards, ITU-T introduced a universally applicable standard called V.90 [7]. It incorporates the features of the K56 and the X2 standards and enables modems supporting V.90 to be universally compatible with the K56 and the X2 standards. The downstream modulation scheme is PCM and the upstream modulation scheme is carried out using V.34 standards. While these asymmetric data rates might not look all that lucrative, they provide the general feeling of a faster communication channel since most users deal with more downstream traffic than upstream traffic.

4. FAX TRANSMISSION

Facsimile (or fax) is defined as the process of sending a document from one terminal to another. In an effort to utilize the existing PSTN infrastructure, traditional fax transmission involved the usage of modems at either communicating terminals. The transmission of a document is preceded by an exchange of capabilities between the sender and the receiver and, at the end of the transmission, a confirmation of delivery sent from the receiver to the sender.

The earlier fax machines, also referred to as group 1 fax machines, were designed to handle fax transmission over an analog telephone network. These conformed to ITU-T T.2 standard for fax transmission and yielded a transmission rate of 6 min per page. With improvements in fax devices, ITU-T later introduced the T.3 standard that allowed for the transmission of up to 3 min per page. Group 3 fax machines, the most widely deployed category of fax machines, were standardized in 1980 for digital facsimile devices to communicate over analog telephone lines. Group 3 machines are based on ITU-T standards T.30 (for fax transmission) and T.4 (for fax file formats) [8] and yield a transmission rate of 6 to 30 seconds per page.

The procedures outlined in the T.30 recommendation comprise of 5 distinct phases: (1) call establishment, (2) control and capabilities exchange, (3) in-message processing and message transmission, (4) postmessage, processing, and (5) call release.

The call establishment phase consists primarily of establishing a connection between the calling and the called terminals and exchanging fax tones. The calling machine dials the telephone number of the called machine and the calling tone (referred to as CNG) is received at the called machine. The CNG tone beeps indicate the existence of a fax call as against a normal voice call. The called fax machine answers the ring signal by going off-hook. After a 1-s delay, the called fax machine sends a 3-s, 2100-Hz tone back to the calling machine.

In the premessage processing phase, the terminals carry out various identification procedures along with command procedures to establish a command set of capabilities for the successful transmission of facsimile data. During the identification phase, the terminals exchange information regarding, among others, the bit rate, page length, data compression format, telephone number, and name of the organization. The called machine sends its digital identification signal (DIS) at 300 bps identifying its capabilities (using V.21 protocol), including its optional features. For example, the called fax machine could send a DIS identifying its capability to support V.17 standard (14,400 bps data rate). On receipt of the DIS, the calling fax machine sends a digital command signal (DCS) locking the called unit into the selected capabilities. The calling machine sends a training check field (TCF) through the modem to ensure that the channel is suitable for transmission at the accepted data rate. The called fax machine sends a "confirmation to receive" (CTR) signal to confirm that the receiving modem is trained (adjusted for low-error operation).

The in-message processing phase takes place in parallel with the message transmission phase since the in-message processing phase handles the signaling required for the transmission of facsimile data. This includes control signals needed for error detection, error correction, and line supervision. The ITU-T T.4 recommendation governs the message transmission phase and addresses issues related to the dimension of the document, the transmission time per scanned line, the coding scheme, modulation/demodulation techniques, and similar. The modem standards that are supported for the transfer of facsimile data include V.27ter (4800/2400 bps), V.29 (9600/7200 bps), V.33 (14,400/12,000 bps), and V.17 (14,400/12,000/9600/7200 bps).

The postmessage processing phase consists of procedures to deal with tasks such as end-of-message signaling, multipage signaling, and confirmation signaling and is used as a precursor to the call release phase. The calling fax machine sends a "return to control" (RTC) command that effectively switches both modems to the 300 bps data rate condition (V.21 standard). The called fax machine sends a message confirmation (MCF) signal indicating the document was received successfully. If multiple pages exist, a multipage signal (MPS) is sent. The partial page signal (PPS) is sent for error correction of the transferred document. In the call release phase, the calling terminal transmits a disconnect (DCN) signal to the called terminal for the release of the call. It is important to note that no response is expected for the call release signal from the called terminal.

5. OTHER MODEMS

The typical bandwidth allocated to each individual user on a telephone line (also referred to as a *subscriber line*) is around 3400 Hz. This places an upper limit on the data rates that a voice band modem can achieve since the transmitting/receiving symbol rate cannot be higher than the available bandwidth. A digital subscriber line (DSL) overcomes this limitation by overlaying a data network onto the existing PSTN. This is accomplished by letting the data network use the same subscriber lines as the POTS (Plain Old Telephone System), with the only exception that the data signals use a different frequency band for data communication.

A device called a POTS splitter is responsible for splitting and recombining the two types of signals: voice signals and data signals, at both ends of the subscriber line. Since the data network and the voice network use the same subscriber line, the telephone companies need only upgrade their switching/terminal devices to handle the data signals. Therefore, the delivery of high-speed data services to customers without considerable investment in infrastructure is possible.

Depending on the data rates that are supported, there are different variations to DSL [9]. ADSL (asymmetric DSL) is characterized by a different data rate from the service provider to the customer (the downstream direction) as compared to the data rate from the customer to the service provider (the upstream direction). The upstream data rates are typically 10 times slower than the downstream data rates and range from 100 to 800 kbps. In sharp contrast, SDSL (symmetric DSL) offers a symmetric data rate; that is, the upstream and downstream data rates are equal. VDSL (very-high-speed DSL) is a new addition to the DSL family and is being developed to provide data rates as high as 25 Mbps in either direction (upstream or downstream).

In ADSL, the transceivers (transmitter and receiver units) are designed to carry more than one logical channel on a single physical channel to support different data rates. In addition, the transceivers support an embedded operations channel (EOC), ADSL overhead channel (AOC) etc that are used mostly for synchronization purposes. The logical data channels on an ADSL link are grouped together as either downstream channels or upstream channels. The downstream channels are simplex in nature and are designated AS0, AS1, AS2, and AS3. Each channel has an allowable data rate up to 8, 4, 3, and 1.5 Mbps, respectively. The duplex channels are named LS0, LS1, and LS3. These support different data rates in upstream and downstream directions and are usually configured as upstream channels by setting the downstream channel data rate equal to 0.

The presence of many different logical channels enables ADSL to support a wide variety of applications since different channels operate at different data rates. However, the total physical bandwidth available in the channel should be greater than the sum of bandwidths of all the logical channels (since a portion of the bandwidth will be consumed by control/synchronization signals).

Since different carrier frequencies are employed for traditional voice transmission, upstream data and downstream data, ADSL uses frequency-division multiplexing (FDM) to multiplex these different signals onto one physical channel. Voice transmission is carried out by letting POTS use the lower-frequency band (0–3400 Hz), while the downstream data is transferred in the higher-frequency band (138 kHz–1.104 MHz). A guard band of approximately 26 kHz is placed between the POTS band and the upstream band to reduce the possibility of interference between voice conversations and data transmission operations.

In sharp contrast to voice band or DSL technologies, cable modems capitalize on the existence of a network of coaxial cables (that are used primarily for video applications) to transmit data. Since the coaxial cables support high bandwidths, cable modems could be used for high-speed data transfers. This is a prime reason for cable technology offering formidable competition to DSL.

Cable systems, however, have traditionally supported only downstream traffic and the available bandwidth is shared among several users since a single line serves many cable subscribers. Therefore, the cable system had to be reconditioned to support upstream traffic from the subscriber back to the coaxial distribution point.

A typical cable system consists of an uplink site that encodes and modulates video content obtained from various sources (tapes, DVDs, live feeds, etc.). The modulated video content is transmitted to downlink site via a satellite. The video content is transmitted from the downlink site to the cable subscribers through a network of coaxial cables. This transfer of video information from the downlink site to the cable subscribers is accomplished with the aid of headend (HE). An HE is responsible for modulating and scrambling each channel that is supported on the cable system and transmitting the scrambled information onto the local hybrid fiber coaxial cable) (HFC) distribution box. Subscribers connect to the HFC through a series of local taps (where each HFC could service 500–2000 homes per fiber node). Data communication is enabled into the cable system by incorporating routing/switching functionality in the HE. This is accomplished by using a broadband router (like a Cisco uBR 7200 series router) for data connectivity. At the customer premise, an additional device would be needed to allow users to utilize the data communication facilities that are available on the same coaxial line that delivers audiovideo content. This additional device is referred to as a *cable modem* (CM).

The downstream communication (from the cable company to the subscriber) is carried out in the frequency range of 54–860 MHz. The upstream communication (from the subscriber to the cable company) is carried out in the frequency range of 5–42 MHz. The manner in which data are transferred over a cable network is specified by DOCSIS (data over cable service interface specification) [10]. Typically, a CMTS (cable modem termination system) is employed at the HE to modulate (or demodulate) the signals sent (or received) from the CM. The CM is associated with the customer premise equipment (CPE) and is responsible for communicating with the CMTS for data communication. CMs support DOCSIS defined connectors, namely, RJ45 ports for Ethernet, RJ11 ports for voice, and F-connector for video.

DOCSIS-compliant devices require the presence of a few servers that provide information regarding IP addresses through the dynamic host configuration protocol (DHCP), time of day timestamps (as defined in RFC 868) and CM configuration files through TFTP (Trivial File Transfer Protocol).

When a CM is powered up, it scans the downstream channel for a synchronizing clock signal. The HE continues broadcasting information regarding upstream channel descriptors, upstream channel frequencies, downstream channel descriptors, and other data that the CM can use for determining the details of the channels it can use

for upstream and downstream communication. Once the CM recognizes the upstream and downstream channels, it begins identifying the bandwidth that can be used for communication.

At this stage, the CM–HFC interface line protocol is considered to be up but the CM is not yet ready to start transferring data with other hosts on the global Internet since it does not yet have an IP address. A DHCP server provides the CM with an IP address, a default gateway, the address of the TFTP server, the CM configuration filename, and the address of a time-of-day (ToD) server that the CM can use for synchronizing its operations. The CM uses the address of the TFTP server to obtain the required files to configure itself as a network entity to handle data transfers across the cable network.

BIOGRAPHIES

Hyuck M. Kwon was born in Korea on May 9, 1953. He received his B.S. and M.S. degrees in electrical engineering (EE) from Seoul National University, Seoul, Korea, in 1978 and 1980, respectively, and his Ph.D. degree in computer, information, and control engineering from the University of Michigan at Ann Arbor, in 1984. From 1985 to 1989 he was with the University of Wisconsin, Milwaukee, as an assistant professor in EE and CS department. From 1989 to 1993 he was with the Lockheed Engineering and Sciences Company, Houston, Taxes, as a principal engineer, working for NASA space shuttle and space station satellite communication systems. Since 1993, he has been with the ECE department, Wichita State University, Kansas, where he is now a full professor. In addition, he held several visiting and consulting positions at communication system industries, and a visiting associate professor position at Texas A&M University, College Station, in 1997. His current research interests are in wireless, CDMA spread spectrum, smart antenna, space-time block code, and MIMO communication systems.

Ravi Bhagavathula received his B.E. degree in electronics and communication engineering in 1997 from Osmania University, Hyderabad, India, and his M.S degree in Electrical Engineering from the Wichita State University in 1998. He is currently working towards a Ph.D. degree in electrical engineering from Wichita State University. He was the recipient of the 2002 Electrical Engineering Outstanding Ph.D. student award. His areas of interest are memory hierarchy design, cache block replacement algorithms, router architectures, layer 3 mobility, and mobile IP extensions using MPLS and VPN architectures.

BIBLIOGRAPHY

1. Recommendation V.42bis, *Data Compression Procedures for Data Circuit Terminating Equipment (DCE) Using Error-Correcting Procedures*, ITU-T *(http://www.itu.int)*.
2. H. Taub and D. L. Schilling, *Principles of Communication Systems*, 2nd ed., McGraw-Hill, New York, 1996.
3. E. Biglieri, D. Divsalar, P. J. Mclane, and M. K. Simon, *Introduction to Trellis-Coded Modulation with Applications*, Macmillan, New York, 1991.
4. Recommendation V.21 (11/88)—*300 Bits per Second Duplex Modem Standardized for Use in the General Switched Telephone Network*, ITU-T *(http://www.int.int)*.
5. Recommendation V.32 (03/93), *A Family of 2-Wire, Duplex Modems Operating at Data Signalling Rates of up to 9600 Bit/s for Use on the General Switched Telephone Network and on Leased Telephone-Type Circuits*, ITU-T *(http://www.int.int)*.
6. Recommendation V.34 (02/98), *A Modem Operating at Data Signalling Rates of up to 33 600 Bit/s for Use on the General Switched Telephone Network and on Leased Point-to-Point 2-Wire Telephone-Type Circuits*, ITU-T *(http://www.itu.int)*.
7. Recommendation V.90 (09/98), *A Digital Modem and Analogue Modem Pair for Use on the Public Switched Telephone Network (PSTN) at Data Signalling Rates of up to 56,000 Bit/s Downstream and up to 33 600 Bit/s Upstream*, ITU-T *(http://www.itu.int)*.
8. Recommendation T.4 (04/99), *Standardization of Group 3 Facsimile Terminals for Document Transmission*, ITU-T *(http://www.itu.int)*.
9. D. J. Raushmayer, *ADSL/VDSL Principles*, Macmillan Technical Publishing, 1999.
10. G. Abe and A. Buckley, *Residential Broadband*, Cisco Press, Indianapolis, 1999.

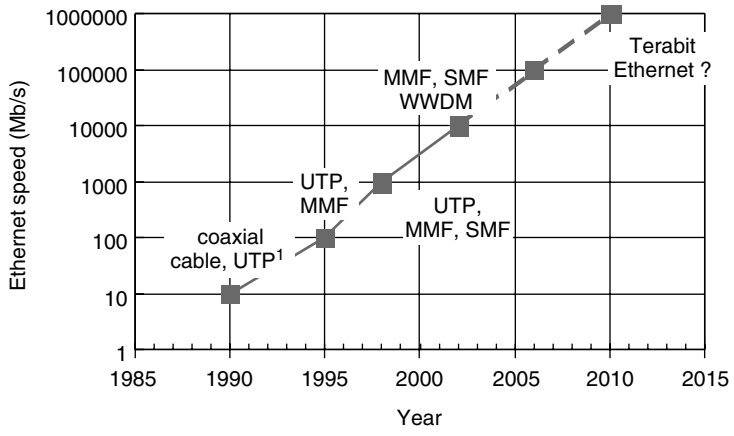
MODERN ETHERNET TECHNOLOGIES

CEDRIC F. LAM
Opvista Inc.
Irvine, California

1. INTRODUCTION

Ethernet was invented in 1973 at Xerox Labs in Palo Alto, California as a medium access control (MAC) protocol for local-area networks (LANs) [1]. Since its invention, Ethernet has gone through many changes in performance, architecture and the underlying technologies. The rapid adoption of the Internet since the early 1990s has made Ethernet the most popular network technology with very fast growth rate. Ethernet has become the ubiquitous means to network servers and desktop computers. Over 85% of the network traffic in today's Internet is generated as Ethernet packets. Not only has Ethernet been enjoying popularity in network computing; it is also becoming more and more popular for internetworking automated manufacturing systems and measurement equipment in factories and research labs. Figure 1 shows the development trend of Ethernet since 1982.

Ethernet has become the most popular network technology among many different competing technologies in the Internet era because of its low cost and simplicity. A 10/100BASE-T Ethernet Network Interface Card (NIC) costs as little as $15 nowadays. The cost for a 1000BASE-T (more commonly known as Gigabit Ethernet) Ethernet port [with 1000 Mbps (megabits per second) throughput] is around $350. Compared to Gigabit Ethernet, a SONET OC-12 intermediate reach interface with 622 Mbps throughput would cost about $3000.



UTP: unshielded twisted pair MMF: multi-mode fiber
SMF: single-mode fiber WWDM: wide wavelength division multiplexing
1. 10 Mb Ethernet using coaxial cables was developed in the 1980s. Ethernet becomes very popular after 10base-T was invented in 1990.

Figure 1. Development trend of Ethernet technology.

2. ETHERNET ARCHITECTURE

IEEE (Institute of Electrical and Electronics Engineers) 802.3 Standard Group charters the development and standardization of the Ethernet technology. The scope of Ethernet covers the physical layer and the data-link layer (layers 1 and 2) of the seven-layer OSI (Open System Interface) reference model. Thus, Ethernet consists of two major layers: (1) the MAC layer, which handles physical-layer-independent medium access control and (2) the PHY (physical) layer, which deals with different physical-layer technologies and various transmission media. In modern Ethernet, the MAC layer and the PHY layer are interconnected with a medium-independent interface (MII). This allows the same MAC design to be used with different transmission technologies. Figure 2 shows the architecture of Ethernet as defined in IEEE802.3 standard [2].

The design of Ethernet itself follows the layered architecture principle. So both the MAC layer and the PHY layer are further divided into sublayers with clearly defined functions.

2.1. Ethernet MAC Frame

Ethernet is a packet-switched technology. Ethernet data are transmitted in packets called *MAC frames*. We may use the terms "frame" and "packet" interchangeably in the following discussions. The format of an Ethernet MAC frame is shown in Fig. 3.

Even though Ethernet has gone through many different generations, the format of the Ethernet MAC frame has never changed. This invariant MAC frame defines the modern Ethernet. By keeping the MAC frame invariant, investment in the upper-layer software can be preserved as the network technology advances. This has tremendous impact on the success of Ethernet technology.

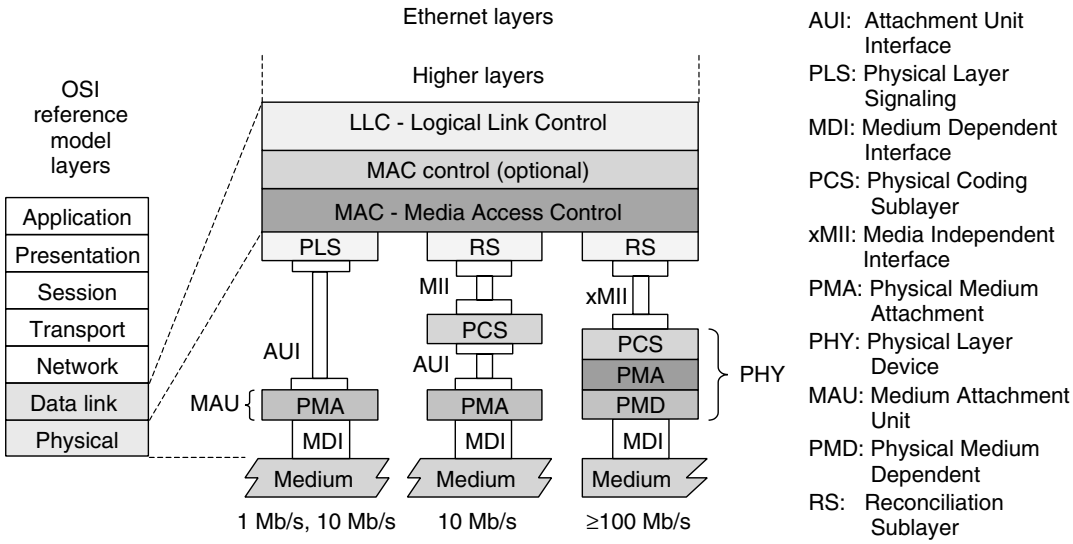


Figure 2. Architecture of Ethernet.

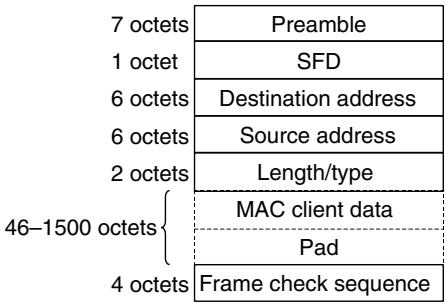


Figure 3. Ethernet MAC frame format.

Ethernet MAC frames have a very simple format with only eight fields (Fig. 3). The leading seven octets in a MAC frame are the preamble field with alternating 0s and 1s for clock recovery purposes at the receiver. This field was useful in early-generation Ethernet, where all the network stations share the same physical channel and data are transmitted in a "bursty" mode. We will see later that the importance of the preamble field diminishes in modern Ethernet, where all stations are joined together with point-to-point dedicated links.

Following the preamble is a one-octet *Start Frame Delimiter* (SLD) field, with the special bit pattern (10101011) to signify the beginning of the actual packet data in the next octet. The next two fields are the Destination and Source Addresses of the MAC frame. Each of them is six octets long. The field following the source address is the two-octet Length/Type field. Ethernet frames are variable length frames with a minimum size of 64 octets and maximum size of 1516 bytes.[1] This field is used to represent the Length of the payload data (from 1 to 1500 bytes) or the type of the MAC frame when its value lies in the range from 1536 to 65535 (2^{16}-1). The payload field follows the Length/Type field and has a size between 46 and 1500 octets. When the actual payload data is smaller than 46 octets, the payload field is padded with zeros to 46 bytes so that a minimum MAC frame of 64-octets is guaranteed.

The last field in the Ethernet MAC frame is a four-octet CRC (Cyclic Redundancy Check) field called frame-check sequence (FCS). Gigabit Ethernet frames may also have a Carrier Extension field following the FCS to extend the size of a packet to a minimum size of 512 bytes, so that the CSMA/CD MAC protocol (which will be explained later) can be supported with a reasonable distance at 1000 Mbps speed.

The simple MAC frame format is the key to the success of Ethernet because it simplifies MAC processing and makes Ethernet devices very cost-effective. On the other hand, Ethernet frames lack the overhead for network management, performance monitoring, fault detection, and localization. This makes large-scale deployment of native Ethernet services a challenging job.

[1] The size of an Ethernet frame does not include the Preamble field and the SFD field.

2.2. Ethernet Address Format

Ethernet uses six-octet-long addresses. The first bit in an Ethernet represents whether the address is a multicast (1) or unicast (0) address. The second bit indicates whether the address is globally administered (0) or locally administered (1). This give a total of 2^{47}-1 globally administered addresses and 2^{47}-1 locally administered addresses.

Ethernet address space is large enough that virtually every Ethernet device in the world can be assigned a globally unique address at the factory. IEEE is in charge of assigning blocks of globally administered addresses to manufacturers of Ethernet interfaces so that each globally administered address is unique in the whole universe. This has significant network implications: (1) there is no need to program the Ethernet MAC address after manufacturing — this reduces the possibilities of human errors; and (2) there is no need for address translation when Ethernet frames are forwarded from one subnetwork to another subnetwork. In some other technologies with very small network address space, the physical addresses of network interfaces in different subnetworks may have the same value. This adds burden to the internetworking devices called bridges (or switches) because address translation must be performed when forwarding data from one subnetwork to another subnetwork, incurring both performance and cost penalties.

2.3. Shared Ethernet: The CSMA/CD Protocol

Ethernet was invented as a medium access control (MAC) protocol for local-area networks (LANs). The first-generation Ethernet adopted a bus architecture with all the network stations sharing a common communication channel as shown in Fig. 4.

Channel arbitration is achieved by the Carrier Sense Multiple Access with Collision Detection (CSMA/CD) protocol [3]. In the CSMA/CD protocol, each station having packets to send first listens to the channel (carrier sense). If the channel is busy, it will wait until the channel becomes idle and clear. If the channel is clear, the station will send the data. However, it is possible for two stations to both sense the channel as clear and try to send data at the same time. In this case, a collision occurs and the stations that are transmitting will stop data transmission and backoff for a random amount of time before re-transmission (collision detection). The stations detecting a collision will also transmit a jamming signal for a certain period of time to ensure that all the stations in the network detect the collision and refrain from transmission.

The CSMA/CD protocol is an unacknowledged protocol; thus, if a station finishes the transmission of a MAC frame before it detects collision, it assumes that that frame is correctly transmitted. In the worst case as illustrated in Fig. 5, station (node) *A* at one end of the bus sends a frame on the bus. Another station, node *B*, at the other end of bus senses the channel as being idle just before the frame from *A* arrives at station *B*. So *A*'s frame will collide with *B*'s frame. By the time the collision propagates back to *A*, a round-trip propagating time has already elapsed since *A* started sending its frame. If *A* finishes transmitting its frame before the

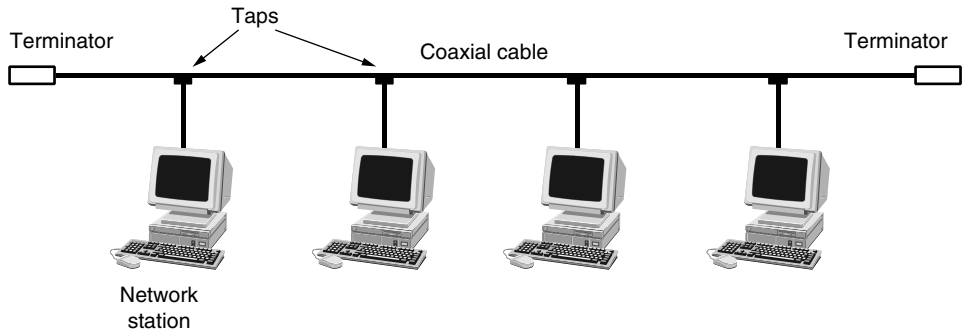


Figure 4. First-generation Ethernet using a shared bus architecture.

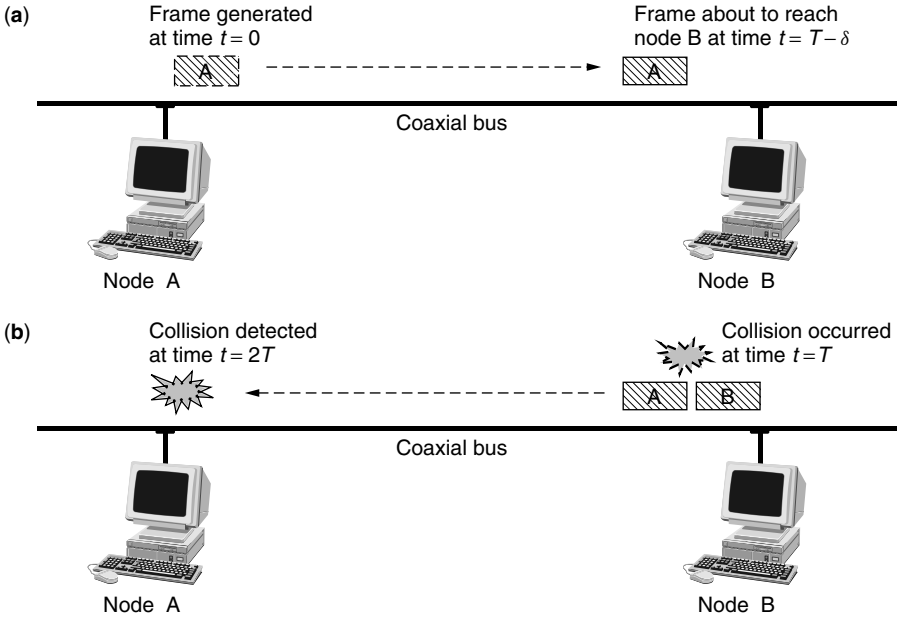


Figure 5. Collisions in a CSMA/CD protocol network.

collision propagates back to A, it will mistakenly think that its transmission was successful. Therefore the minimum packet length, the maximum network size and the transmission speeds are tightly coupled in the CSMA/CD protocol. Ethernet has chosen a fixed minimum packet size of 64 octets. As can be expected, as the transmission speed increases, the network size has to scale inversely for the CSMA/CD protocol to function properly. Thus the CSMA/CD-protocol-limited transmission distance is 2500 m and 250 m for 10-Mbps and 100-Mbps Ethernet, respectively. For Gigabit Ethernet, this would become a very limiting distance of only 25 m. In order preserve the CSMA/CD protocol for Gigabit Ethernet, IEEE 802.3 standard defined the Carrier Extension operation and Frame Bursting techniques to extend the protocol-limited transmission distance and retain the bandwidth efficiency. Since virtually no Ethernet device is implemented with Carrier Extension and Frame Bursting, we will not discuss it here. It should also be noted that Ethernet transmission distance is also limited by the physical technology such as available transmission power, receiver sensitivity, and signal attenuation and degradation as data propagates in the channel.

The CSMA/CD protocol is also called *half-duplex operation* as a station cannot transmit and receive at the same time. All the stations sharing the same bandwidth form a collision domain. Since only one station in a collision domain can be transmitting at a time, it can be expected that the average network performance degrades as the number of stations in a collision domain increases.

2.4. Repeaters

As explained before, Ethernet transmission distance is limited not only by the CSMA/CD protocol but also by signal degradation in the channel. For example, the collision domain size for 10-Mbps Ethernet is 2500 m. However, the physical limit for 10BASE-5 thick coaxial cable Ethernet is only 500 m. In order to extend the transmission distance, a repeater is required [2].

Repeaters usually have multiple ports. A repeater receives the signal from one port, recovers the MAC frame, and retransmits (broadcasts) the frame to every other port. If a repeater detects simultaneous transmission at more than one port, it will transmit a jamming signal to all the ports to cause collision detection by every workstation. Therefore, all the stations joined by a repeater belong to the same collision domain. Repeaters introduce signal delays. Such delays need to be taken into account when calculating the collision domain size. Because repeaters terminate the data at the MAC interface layer, it also enables Ethernet with different physical media (e.g., coaxial cable and optical fiber) to be interconnected. However, repeaters cannot be used to interconnect Ethernet with different speeds. "Bridges" (also called "switches") are needed in that case.

2.5. Modern Ethernet — Hubbed Architecture

In the late 1980s and early 1990s, structured wiring of category 5 unshielded twisted pairs became popular. In structured wirings, all the connections terminate at a central location (usually in a building wiring closet). Hubbed Ethernet was introduced to take the advantage of structured wiring.

In hubbed Ethernet, all the stations are connected to a hub (usually located at a wiring closet in a building) through point-to-point connections as shown in Fig. 6. There is no direct station-to-station communication. All the transmissions between stations have to go through the hub. Compared to the coaxial bus architecture, the hubbed topology has several advantages. First, the physical hub provides a convenient location where all the network connections can be centrally managed. In the bus architecture, the network connection is disrupted during the addition and removal of a workstation from the bus as the coaxial cable must always be properly terminated to prevent signal reflection. In the hubbed architecture, each port is individually terminated within the hub. Also, offending stations can be easily isolated from the network by disabling the corresponding hub port that the station is attached to, making the network more reliable.

It should be realized that there are two types of hubs used in Ethernet: a repeater hub and a switch hub. A repeater hub runs the CSMA/CD protocol. It allows only one station to transmit at a time in half-duplex mode. Repeater hubs are less costly and more commonly used for 10-Mbps (10BASE-T) Ethernet. Switched hubs with higher performance are more popular in modern Ethernet. More 100-Mbps Ethernet devices are running in the switched full-duplex mode than the half-duplex mode. Although the CSMA/CD protocol has been defined for Gigabit Ethernet, there is no Gigabit Ethernet devices built using half-duplex mode. Half-duplex operation is not defined in the 10-Gbps Ethernet standard, which will be finalized during the year of this writing (2002).

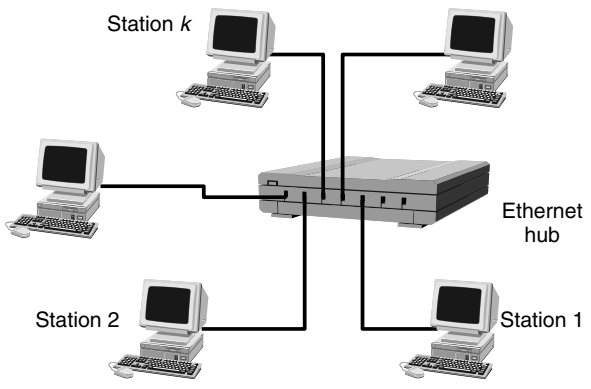


Figure 6. Hub Ethernet physical topology.

In full-duplex operation, the hub switch receives packets from stations attached to its ports and switches the packets to the appropriate output ports according to the destination address and an address table stored in the switch. Since all the connections are dedicated point-to-point links between switch ports and end stations, there is no multiple access and no need for the carrier sense operation. Of course, there is no collision to detect either. Furthermore, there is a separate transmitting path and a separate receiving path. This enables simultaneous transmission and reception, namely, full-duplex operation mode. It should also be realized that in a purely switched Ethernet environment, the transmission distance is limited only by physical signal impairments during transmission.

3. ETHERNET SWITCHES

Switches were initially called "bridges." Early bridges were mostly implemented in software. Bridges became switches when their functions were implemented in hardware, as a result of the development of silicon integrated-circuit technology.

A bridge is an internetworking device connecting different subnetworks [4]. Bridges enable networks with different speed, media, or even different protocols to be connected together. A bridge connecting different protocol subnetworks also needs to perform format translation and address mapping. We only cover Ethernet switches here. Ethernet switches are layer 2 devices switching at the level of Ethernet frames. Switches can be used to join networks running the CSMA/CD protocol. The use of switches enables the network to reach beyond the limit of collision domain size. In a shared Ethernet environment with a large number of stations, switches can also be used to improve the network performance by segregating the network into multiple interconnected collision domains with a smaller number of stations.

Ethernet switch operations and maintenance are defined in the IEEE 802.1D Standard [5]. Figure 7 shows the functional block diagram of an Ethernet switch. Switches have two major functions: packet forwarding and filtering. A switch interface to a workstation is called a port. Switch ports work in a "promiscuous" mode. A port examines all the input frames for the Destination Addresses. A switch uses a source address table (SAT) to determine whether the packet should be forwarded or filtered. The SAT stores MAC addresses and the associated switch ports. Figure 8 illustrates the operation of an Ethernet switch. If the Destination Address is found in

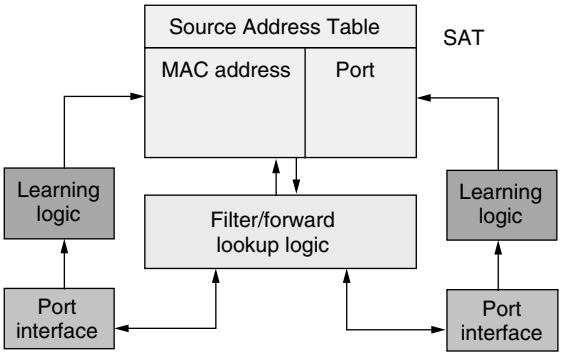


Figure 7. Functional block diagram of an Ethernet switch.

the SAT and its associated port is the same port where the packet arrives from, then the source node and destination node are attached to the same port of the switch and the packet is not forwarded; in other words, the arriving packet is filtered. If the destination node address is associated with another port in the SAT, then the packet is forwarded to that particular port. In the case when the destination node address is not in the SAT or if the packet is a multicasting packet, that packet is flooded to all the ports except the one it arrives from.

Switches use a switching fabric to route packets between ports. To achieve good performance, the switching fabric should have a capacity equal to the aggregate bandwidth of all the switch ports.

There are two ways that Ethernet switches use to populate the SAT. Static entries are entered through the management system and do not expire until updated by the system administrators. Dynamic entries are acquired through backward learning. In backward learning, switches examines the Source Address field of the arriving packets. If that address is not in the SAT, then it is entered into the SAT and associated with the arriving port. These addresses will time out if they are not active for a certain amount of time (300 s default value in IEEE 802.1D Standard). This allows the attached Ethernet interfaces to be moved from one location to another. Moreover, new entries will replace old entries when the SAT becomes full. To achieve good performance, the SAT size should be comparable to the expected number of stations connected to the switch to reduce the volume of broadcast traffic in the network.

4. ETHERNET PHYSICAL LAYER

4.1. Transmission Medium

Ethernet has gone through many different generations. Different physical (PHY)-layer technologies have been invented for different transmission media. The first-generation Ethernet (10BASE-5) uses thick coaxial cable as the transmission medium. Thick coaxial cables enable a transmission distance of 500 m. However, these cables are very inflexible and are used mostly in building risers. An *attachment unit interface* (AUI) has been devised to enable the analog transceiver called *medium access unit* (MAU) to be separated from the MAC digital processing unit (Fig. 2) and connected through a 50-m-long flexible cable with 15-pin IEC60807-2 connectors. As technology improves, the PHY layer and MAC layer become integrated into one circuit board and more flexible transmission media have been adopted. 10BASE-2 Ethernet using thin coaxial cable has a transmission distance of 185 m and unshielded twisted-pair (UTP) interfaces (10/100/1000BASE-T) with 100 m transmission distance have been subsequently introduced. A standard interface between the PHY layer and the MAC layer called *media-independent interface* (xMII) is adopted in modern Ethernet to separate the MAC design and the PHY design.

Category 5 UTP cables are by far the most popular medium for Ethernet connections to desktop computers for data rate up to 100 Mbps. Category 5 cables have the advantage of low cost and easy installation. The use of category 5 cable was enabled by modern signal processing techniques and silicon technology development.

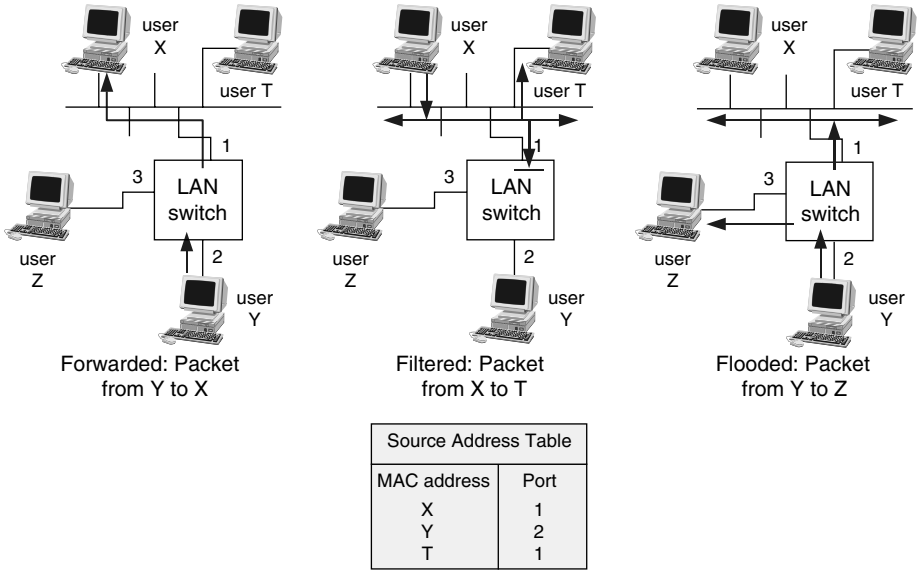


Figure 8. Ethernet switch operation.

As the transmission speed and transmission distance increased, the Ethernet community also moved from copper-based media to optical fiber–based media. Optical fiber has the advantages of virtually unlimited bandwidth and very low loss. An optical light source such as an LED (light-emitting diode) or laser diode is used as the transmitter and a photodetector is used as the receiver. Like copper cables, there are different kinds of optical fibers. Multimode fibers (MMFs) were commonly used in the past for short-distance communications. These fibers have large core diameters (55 μm or 62.5 μm are the two most commonly used). The large core diameter makes coupling light into the fiber an easier job compared to coupling light into single-mode fibers. However, in an MMF, optical signals can propagate in many modes with different speeds. This is called *modal dispersion*. Since light detection is mainly intensity-based, the received light pulses will get smeared after transmission in a multimode fiber. Modal dispersion limits the transmission distance as well as data speed.

Modern lightwave communication systems are increasingly using single-mode fiber (SMF). Standard SMF has a core diameter of 10 μm and therefore requires careful handling. As its name implies, optical signals can only propagate in one mode in an SMF. Therefore, SMF allows signals to propagate a longer distance. SMF suffer from chromatic dispersion. In a digital communication system, light signals are transmitted as pulses and have a finite frequency spectrum. Different frequencies of light travel at different speeds in an SMF. This incurs pulse broadening and limits transmission rate and transmission distance. Chromatic dispersion is a less severe effect compared to modal dispersion. It can be readily compensated if necessary (at a certain cost, of course). Single mode fiber has been adopted as one of the transmission media for Gigabit Ethernet. As the speed of Ethernet is increased to 10 Gbps, coarse wavelength division multiplexing (WDM) has also been adopted in one of the PHY designs. In the coarse WDM PHY design (called "10GBASE-LX4" in Ethernet Standard), four wavelengths separated by 20 nm in optical spectrum are used to carry the 10-Gbps payload in a parallel fashion. This reduces the requirement for high-speed electronics. The four wavelengths are combined into and separated from the same fiber using passive wavelength-division multiplexers.

Tables 1 and 2 specify the reach of Gigabit and 10 Gigabig Ethernets in different fiber media and the reach with different wavelength transmitters as specified in the IEEE 802.3 Standard. Signal attenuation and dispersion in optical fiber depends on the transmitter wavelength.

4.2. PHY Sublayers

4.2.1. MII and RS Sublayers. In modern Ethernet, the MAC layer and PHY layer are separated by the media-independent interface (MII). The MII uses parallel connections for control, timing, and data signals to reduce the digital processing speed requirements. The acronym MII is actually used for 100-Mbps Ethernet. For Gigabit and 10-Gigabit Ethernet, this interface is called "GMII" and "XGMII," respectively. Data are transmitted in units of 4 bits (called "nibbles") in MII, 8 bits (called "octet") in GMII, and 32 bits in XGMII.

The "reconciliation sublayer" (RS) is an abstract layer that defines the mapping of protocol primitives between the MAC layer and the PHY layer to the MII signal pins.

4.2.2. Physical Coding Sublayer (PCS). The PCS sublayer is responsible for line-coding the signals. There are

Table 1. Cable Length Specifications for 1000base-SX (850 nm Short-Wavelength Transmitter) and 100base-LX (1300 nm, Long-Wavelength Transmitter) Ethernet Interfaces

	1000base-SX		1000base-LX	
Fiber Type	Modal Bandwidth at 850 nm (MHz·km)	Range (ms)	Modal Bandwidth at 1300 nm (MHz·km)	Range (ms)
62.5-μm MMF	160	2–220	—	—
62.5-μm MMF	200	2–275	500	2–550
50-μm MMF	400	2–500	400	2–550
50-μm MMF	500	2–550	500	2–550
10-μm SMF	N/A	Not supported	N/A	2–5000

Table 2. Cable Length Specification for 10Gbase Ethernet Interfaces

	62.5 μm MMF		50 μm MMF			10 μm SMF	
Wavelength (nm)	850	850	850	850	850	1310	1550
Modal bandwidth (min; overfilled launch) (MHz · km)	160	200	400	500	2000	N/A	N/A
Operating distance	28 m	35 m	69 m	86 m	300 m	10 km	40 km
Channel insertion loss	1.61	1.63	1.75	1.81	2.55	6.5	13.0
Dispersion (ps/nm)	—	—	—	—	—	—	728

several reasons for line coding: (1) line coding provides enough transitions in the bit-stream for the receiver to recover signal clocks, (2) line coding provides redundant symbols for certain physical layer signaling purposes, and (3) line coding may encode multiple binary digits into a single transmitted symbol to reduce the transmission bandwidth requirement. This is especially important for the bandwidth-limited copper medium.

In 10-Mbps Ethernet, Manchester coding was used to embed the clock signal into the transmission data. Zeros and ones are respectively represented by high-low and low-high transitions in Manchester coding. Clock recovery is very easy. However, twice the bandwidth is required to transmit the data so that 20 MHz bandwidth is required for 10-Mbps signals. In 10-Mbps Ethernet with bus architecture, all the transceivers share the same physical bus, so transmission is bursty and a preamble in front of each packet frame is required for clock recovery at the receiver.

When Ethernet moved to the point-to-point architecture, a dedicated path is established between each hub port and the Ethernet station, and, hence, there is actually continuous physical signaling between each point-to-point transceiver pair. Burst-mode receiver operation is not required anymore. Even though the transmitted data may be bursty, idle periods between data packets are filled with idle symbols. Although the preamble has been defined for Ethernet frames, their importance has diminished in the point-to-point architecture except for backward compatibility.

Ethernet has been designed to operate on different media using different technologies. For example, 100-Mbps Ethernet has been designed to operate on both Category 3 and Category 5 cables and MMF. The most commonly used 100-Mbps Ethernet (100BASE-TX) uses two pairs of Category 5 cables and a line coding scheme called "4B/5B." The 4B/5B coding encodes 4 data bits into 5 bits and has also been used in FDDI (fiber distributed digital interface). 100base-T4 and 100base-T2 are defined for four pairs and two pairs of Category 3 cables, respectively. Since Category 3 cables have worse frequency responses than Category 5 cables, bandwidth efficient line coding schemes are used. In 100BASE-T4, an encoding scheme called "8B6T," which encodes eight binary digits into six ternary symbols, is used. 100BASE-T2 uses an encoding scheme called "PAM5 $\times$ 5." The transmitted data are encoded into two sets of five-level pulse-amplitude-modulated (PAM) symbols on two pairs of Category 3 cables. The resultant symbol rates of both 8B6T and PAM5$\times$5 encoding schemes are 25 Mbaud.

Gigabit Ethernet uses an 8B/10B encoding scheme, which encodes 8 binary digits into 10 bits. In the 8B/10B encoding scheme, there are no more than four continuous zeros or ones. The 8B/10B encoding scheme selects 256 patterns out of the 1024 possible 10-bit codes to represents the 8 data bits. Some of the extra codewords are used to represent idle symbols, and control sequences such as start of data and error conditions.

8B/10B encoding is very popular in high-speed data transmissions. It is also used for Fiber Channel systems invented by IBM. The price paid for 8B/10B encoding is a 25% overhead. So to transmit 1000 Mbps data, the physical layer needs to handle 1250 Msps (million symbols per second).

As Ethernet speed increases to 10 Gbps, the 25% overhead introduced by 8B/10B encoding makes the design of high-speed transceivers difficult. Therefore, 10-Gbps Ethernet adopted a new encoding scheme called "64B/66B," where 64 binary digits are encoded into 66 bits, with a 3% overhead. So instead of using 12.5-Gbps optical transceivers, 10.3-Gbps transceivers will suffice.

4.3. Physical Medium Attachment (PMA) Sublayer

The physical medium attachment sublayer transforms the PCS output codes into the physical symbols to be transmitted on the actual medium in the transmit path and performs the reverse operation in the receive path. For example, in Gigabit Ethernet, the PMA takes the 8B/10B-encoded PCS output from a parallel interface and transforms it into serial bit symbols to be transmitted on the actual fiber medium. In the receive path, it converts the received serial bits into 8B/10B codewords with parallel output. Therefore, the PMA layer is also commonly called "SERDES (SERializer-DESerializer)" in Gigabit and 10-Gigabit Ethernets.

4.4. Physical-Medium-Dependent (PMD) Sublayer

The physical-medium-dependent sublayer specifies the electrical and/or optical characteristics of the actual transceivers. These include properties such as output signal power, optical wavelength, modulation depth, receiver sensitivity, and saturation power. Table 3 shows typical characteristics of two different PMDs for Gigabit Ethernet. The 1000base-SX standard PMD uses short-wavelength (860-nm) lasers as transmitters and 1000base-LX standard PMD uses long-wavelength (1300-nm) lasers for transmission.

4.5. Medium-Dependent Interface (MDI)

The *medium-dependent interface* (MDI) specifies the electrical or fiber connectors used to connect Ethernet devices. For 10BASE-2 thin coaxial cable Ethernet, the BNC connector has been adopted. The most commonly seen MDI is the RJ45 connector for UTP cables. The fiber connector specified for Ethernet is the SC connector. Figure 9 shows some commonly used Ethernet media and their associated connectors (i.e., MDI).

5. 10-GIGABIT ETHERNET

We devote a section to 10-Gigabit Ethernet because it is the latest Ethernet standard. In fact, at the time of this writing, the 10-Gigabit Ethernet standard is still being finalized by the IEEE 802.3ae Working Group even though the standard has already been quite stable after many iterations of revisions [6].

Figure 10 shows the architecture of 10-Gigabit Ethernet. As shown in the figure, there are three types of 10-Gigabit Ethernet. 10 Gb/s transmission technology is still the state-of-the-art fiberoptic technology and is used

Table 3. Transmitter and Receiver Characteristics of 1000base-SX and 1000base LX

		1000base-SX	1000base-LX	
Medium:		MMF (50, 62.5 μm)	MMF (50, 62.5 μm)	SMF (10 μm)
Wavelength (λ):		770–860 nm	1270–1355 nm	
Transmitter	Spectral width	0.85 nm	4 nm	
	T_{rise}/T_{fall} (max: 20–80%)	λ > 830 nm: 0.26 ns; λ ≤ 830 nm: 0.23 ns	0.26 ns	
	Average launch power (max)	Lesser of class I safety limits or maximum receive power	−3 dBm	
	Average launch power (min)	−9.5	−11.5 dBm	−11 dBm
	Extinction ratio (min)	9 dB	9 dB	
	RIN (max)	−117 dB/Hz	−120 dB/Hz	
Receiver	Average receive power (max)	0 dBm	−3 dBm	
	Average receive power (min)	−17 dBm	−19 dBm	
	Return loss (min)	12 dB	12 dB	

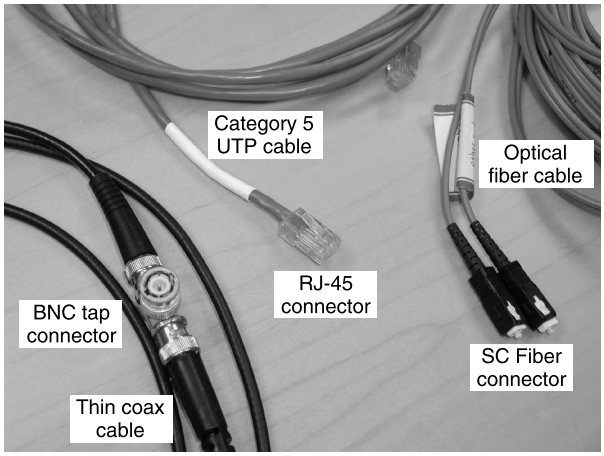


Figure 9. Some commonly seen Ethernet media and connectors (MDI).

mostly in wide-area backbone networks. In fact, 10-Gigabit Ethernet is targeted toward metropolitan-scale backbone network applications. Traditionally, wide-area networks (WANs) are dominated by SONET (Synchronous Optical NETwork) technology [7]. The 10Gbase-W standard is also called "WAN PHY." It includes a WAN Interface Sublayer (WIS) to encapsulate Ethernet frames into frames compatible with SONET Synchronous Payload Envelope (SPE). SONET framing includes extensive overhead bytes for operation, maintenance, and alarm signaling and performance monitoring. It should be noted that not all the SONET overhead fields are implemented by WIS. It should also be noted that the WAN PHY does not define SONET-compatible electrical and optical output, which is very stringent. The "SONET-Lite" framing introduced by the WAN PHY only makes it easier to map Ethernet traffic onto SONET equipment.

Both the 10Gbase-W and 10Gbase-R (Fig. 10) standards specify 64B/66B encoding in the PCS sublayer. Compared to the 8B/10B encoding used in Gigabit Ethernet, the overhead is reduced from 25% to 3%. This makes it easier to implement the high-speed optoelectronic front end.

The 10Gbase-X standard, however, continues to use the 8B/10B encoding scheme. The only PHY defined for 10Gbase-X is the 10Gbase-LX4 standard, which uses four coarse WDM wavelengths as shown in Fig. 11. Each wavelength is carrying a stream of symbols at 3.125 Gbaud/s speed.

10-Gigabit Ethernet will support both MMF and SMF fibers. SMF and 1.5-μm lasers will be used for connections up to 40 km without amplification. The 1.5-μm-wavelength signals have the lowest attenuation in optical fibers. They can also be easily amplified by mature Erbium-Doped Fiber Amplifiers (EDFAs) [8]. In fact, long-haul dense WDM (DWDM) system are designed mostly to operate around the 1.5 μm wavelengths. It should be noted that the 40-km transmission distance is limited by the low-cost transceivers specified in the IEEE 802.3 standard. There are a lot of commercially available nonstandard systems that enable native Gigabit and 10-Gigabit Ethernet signals to go beyond the distances specified in the IEEE 802.3 standard. Transmission of Gigabit Ethernet signals over 1000 km has been demonstrated in the MONET (Multiwavelength Optical NETwork) project that DARPA (Defense Advanced Research Program Agency) funded [9].

6. ETHERNET FOR THE FIRST MILE (EFM)

Ethernet in the first mile (EFM) is a new effort launched by the IEEE 802.3 committee in 2000 to work on the standards and technology required to provide Ethernet services by telecom service providers. The EFM study

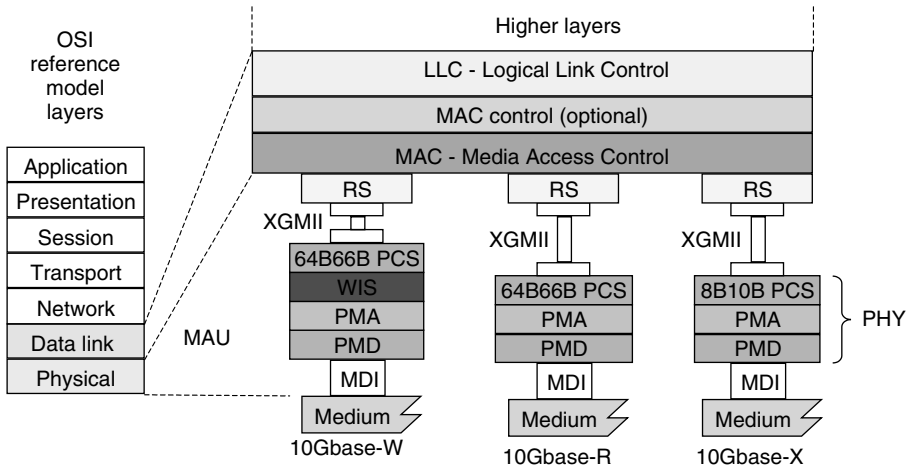
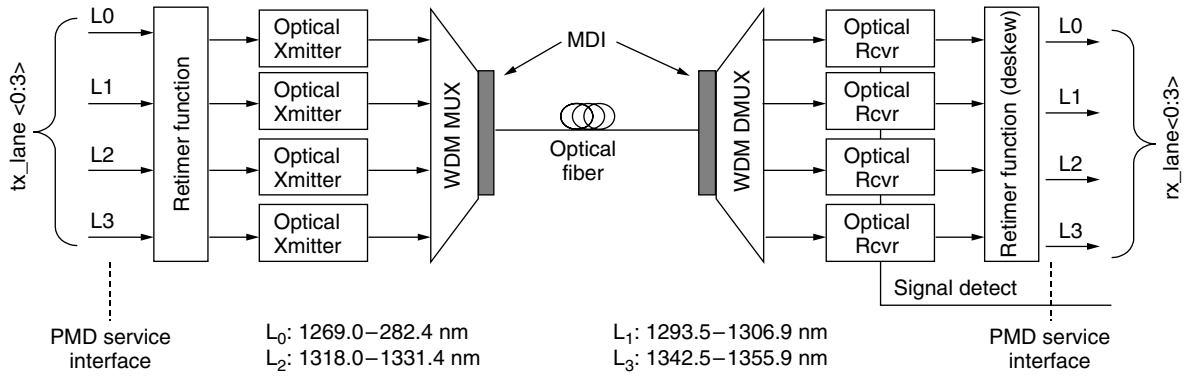


Figure 10. Architecture of 10-Gbps Ethernet.



Parameter	62.5 μm MMF	50 mm MMF		10 μm SMF	Unit
Modal bandwidth at 1300 nm (overfilled launch)	500	400	500	n/a	MHz•km
Link power budget	8.0	8.0	8.0	9.0	dB
Operating distance	300	240	300	10,000	m

Figure 11. 10Gbase-LX4 PMD.

group is registered as IEEE 802.3ah. The scope of the IEEE802.3ah is illustrated in Fig. 12. It covers the operation of Ethernet on a copper medium with extended reach and operation temperature, point-to-point Ethernet operation over single fiber, Ethernet passive optical network (EPON) and OAM (operation, administration, and management) issues and requirements for providing Ethernet services.

6.1. Copper PHY with Extended Reach and Operation Temperature

Ethernet has been defined to operate on Category 5 unshielded twisted pairs (UTP) up to 1000 Mbps and 100 m from a station to the hub. This group is studying the operation of Ethernet on a copper medium with longer reach, extended temperature range (for outdoor applications), and on lower-grade copper pairs. Advanced signal processing and error correction are implemented.

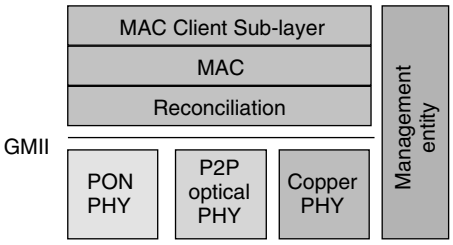


Figure 12. Scope of IEEE802.3ah (EFM) study group.

6.2. Single-Fiber Point-to-Point Operation

For this type of service, 100-Mbps Ethernet (100BASE-FX) has been defined to operate on MMF with a reach of 2 km. Gigabit Ethernet (1000base-LX) has been defined to operate on SMF with a reach of 5 km using 1300-nm lasers. For all the defined fiber-type Ethernet PHYs, full-duplex fiber pairs have been used, with one fiber for transmission and one fiber for reception. One job of the IEEE 802.3ah

study group is to define point-to-point (switched) Ethernet operation over a single fiber with extended reach at 100 and 1000 Mbps speeds. Fiber has the advantage of low loss and virtually unlimited bandwidth. However, fiber termination and connections are still more expensive compared to copper pairs. Therefore, single-fiber operation with longer reach has the advantage of saving cost and enabling service providers to cover a large area from a single central office. Wavelength-division duplex using widely separated 1.3/1.5 μm wavelengths and coarse WDM filters (called diplexers because they split and combine the 1.3 and 1.5 μm wavelengths) is the most popular method to achieve full-duplex operation in a single fiber.

6.3. Ethernet Passive Optical Network (EPON)

Passive optical networks (PONs) have been proposed as an economical way to provide future-proof broadband services. The architecture of an Ethernet PON is shown in Fig. 13. A PON consists of three major parts: an optical line termination (OLT) unit at a service provider's central office (CO), the distribution fiber plant itself, and an optical network unit (ONU) at each customer premise. PON is a distribution network. The signal from the OLT is distributed to ONUs at a remote node (RN). The RN can be either a power splitter (as in traditional PONs) or a wavelength router (as in WDM PONs). In both cases, the RN is a passive element. The term *passive optical networks* comes from the fact that the distribution fiber plant between the CO and the customer premise is passive, that is, there is no electrical power required in the field. This not only reduces operation cost but also improves the reliability.

Here we focus only on PONs using power splitters as the distribution mechanism. The OLT is basically an Ethernet switch that routes Ethernet packets between ONUs and also forms the gateway between the PON and the backbone network. Usually, the ONU outputs consists of T1 interface for legacy voice and data connections, PSTN (plain switched telephone network) interface for telephone services, and 10/100BASE-T interface for data applications.

6.3.1. Physical Design Considerations. The goal of EPON is to achieve transmission of Gigabit Ethernet with a distance of at least 10 km between the OLT and ONU, and a splitting ratio of 1:16 or higher. The available transmitter output power and receiver sensitivity will impose a limit on the transmission distance, remote node splitting ratio, and transmission speed. As an example, the ITU Standards (ITU Rec. G.983.1, G.983.2, G.983.3) [10] for an ATM PON (which carries ATM cells as opposed to Ethernet frames) specify 32-way split with 20-km transmission distances for an aggregate transmission speed of 622 Mbps (OC-12) in the fiber.

In a typical PON system, the downstream (from CO to customer premise) and the upstream (from customer premise to CO) signals are multiplexed on the same fiber using 1.5 and 1.3 μm wavelengths to avoid interference. Standard SMF has zero dispersion at 1.3 μm. This makes it possible to use low cost 1.3-μm-wavelength Fabry–Perot lasers for transmission without worrying about multimode output and chromatic dispersion. The non–zero dispersion at 1.5 μm wavelength will induce pulse broadening and hence will result in performance penalty. To support Gbps bit rate and beyond, DFBs (distributed feedback) lasers with single-mode output must be used at 1.5 μm wavelength. DFB lasers are more difficult to manufacture and therefore more expensive. Since the downstream laser is shared among all the ONUs, 1.5-μm lasers are used for downstream transmission, while 1.3-μm lasers are used at the more cost-sensitive ONUs.

6.3.2. Point-to-Multipoint Operation. In a PON system, the downstream signal is broadcast to all ONUs using the passive splitter. Each ONU detects its own packets by examining the destination address field as in the CSMA/CD protocol. The upstream transmission is achieved in a fashion similar to TDM. Time is divided into units of time quanta. Each ONU is granted permission to transmit at certain time instants for a specific number of time quantas by the OLT. The OLT may dynamically change the grant to transmit according to the network load and service-level agreement. This is called *Dynamic Bandwidth Allocation* (DBA).

Since each ONU is at a different distance from the OLT, it is necessary to align the logical time reference from each ONU to the OLT in order to avoid collision of upstream packets at the remote node. This is achieved through a ranging process illustrated in Fig. 14. The OLT periodically sends out ranging grant frames (or sync frames) to ONUs. An ONU will listen to the ranging grants when powered on. On receiving a ranging grant, it will send back a ranging request frame. The OLT will calculate the round-trip time between the OLT and ONU from the delay between the ranging grant and request frames. This information is sent back to the ONU so that the ONU can adjust its timing reference. If more than one ONU is trying to range at the same time, a backoff mechanism similar to the CSMA/CD protocol can be used to resolve collision.

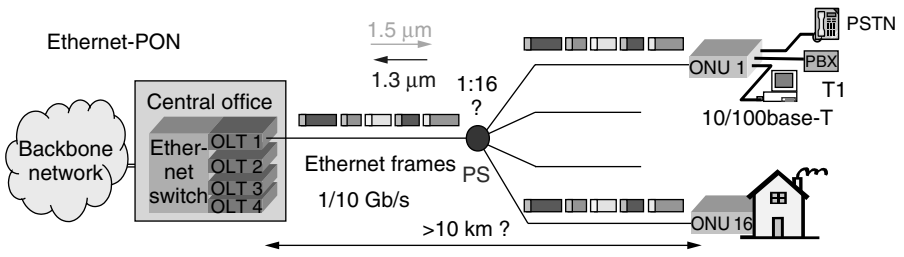


Figure 13. Ethernet PON architecture. A passive power splitter (PS) is used as the remote node.

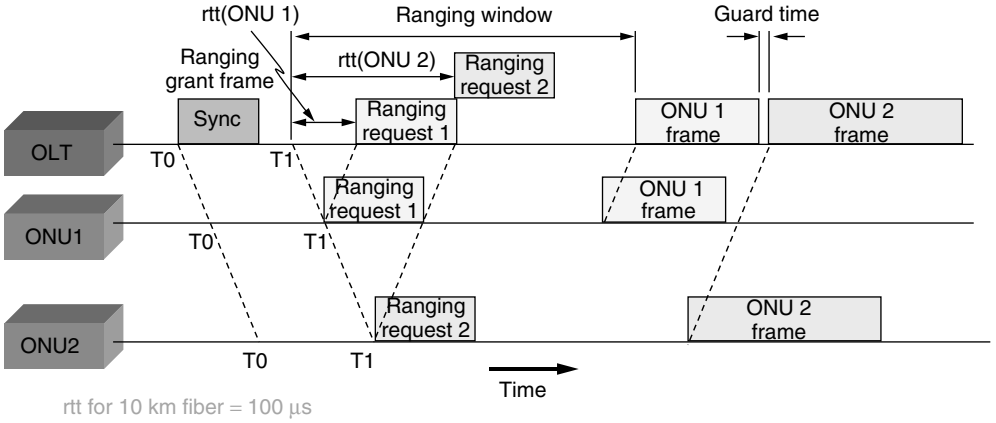


Figure 14. Ranging in a PON system.

We have seen that modern Ethernet assumes point-to-point operation between the hub and each station. The proper routing of Ethernet packets by the switches rely on this point-to-point architecture. However, a PON system is really a point-to-multipoint system. Therefore, although all the ONUs are on the same side of the OLT (which functions like a switch), if an ONU broadcasts a packet to other ONUs, this packet has to be relayed by the OLT. This also applies to other inter-ONU communications. The results of the point-to-multipoint nature of a PON system is that a point-to-point (P2P) emulation layer needs to be added to the OLT to function as the bridge between ONUs.

Previously, we also described that in modern Ethernet with point-to-point operation there is a continuous clock signal between two stations. Idle symbols are added by the PCS sublayer to keep the transmitter and receiver in sync when there are no data to transmit. This does not apply to PON systems. In a PON system, the downstream transmission from the OLT to ONUs is continuous. However, the reverse upstream transmission is bursty. A burst-mode receiver is required at the OLT and should have the capability to quickly synchronize with the transmitter with minimum preamble size. It should also be able to automatically adjust the threshold level for digital demodulation, as the received power from ONUs at different distances from the OLT will be different. At Gbps speed, these requirements are not easily achieved.

6.3.3. Other EPON Issues. There are other issues relating to OAMs covered by the EPON study group. Examples are loopback function to allow for link fault localization and detection and physical-layer device management. In fact, Ethernet PON is still an area of active research. Although there are proprietary prototypes developed by some system vendors, the IEEE 802.3ah study group still has a lot of work to do before a draft standard becomes available.

7. CONCLUSION

Ethernet has become the most popular technology for layer 2 transport. It has changed from a protocol for desktop LAN application to a transport technology for both LAN and MAN (metropolitan-area network) applications.

Some people believe that 10-Gbps Ethernet will also become a significant technology for long-haul wide-area network (WAN) applications. Traditionally, the high-bandwidth and long-distance backbone connections have been served using the SONET technology. The removal of the CSMA/CD protocol in high-speed Ethernet has enabled Ethernet to go long distances. Gigabit products are now widely available at much more affordable prices than SONET and ATM products [11] with comparable performance. Conventional SONET equipment has a very rigid time-division multiplexing (TDM) hierarchy. To compete with Ethernet, next-generation SONET boxes will implement TDM with statistical multiplexing to gain efficiency, and lightweight SONET functions to make them much more cost-effective.

The world of telecommunications is moving into packet switching, and Ethernet frames are really the dominating format in data communications. SONET and ATM equipment being designed for circuit switching does not provide an efficient platform for carrying the ubiquitous Ethernet traffic. Besides the cost and overhead efficiency, another advantage of end-to-end native Ethernet service is to simplify network management because there are no multiple platforms to manage. It should be noted that there are more people in traditional telecom companies who understand SONET better than Ethernet, and the reverse is true for data(communication) companies. The extensive SONET OAM overhead bytes are important for service providers to manage their systems. Ethernet, on the other hand, does not have many management capabilities because of its simplicity and cost-effectiveness. In order to provide end-to-end native Ethernet services, management functions need to be added to Ethernet technology. Standard bodies such as IEEE and IETF (Internet Engineering Task Force) are busy working on these issues. The important trend is that as data become the dominating network traffic, Ethernet will become the ubiquitous technology for both access and transport networks in LAN, MAN, and WAN environments.

8. USEFUL WEBSITES

1. IEEE 802 LAN/MAN Standard Committee, *http://www.ieee802.org/*.
2. Search for IEEE standards, *http://ieeexplore.ieee.org/lpdocs/epic03/standards.htm*.
3. IEEE 802.3 CSMA/CD (ETHERNET), *http://www.ieee802.org/3/*.
4. IEEE P802.3ae 10 Gb/s Ethernet Task Force, *http://grouper.ieee.org/groups/802/3/ae/index.html*.
5. Gigabit Ethernet Alliance, *http://www.gigabit-ethernet.org/*.
6. 10-Gigabit Ethernet Alliance, *http://www.10gea.org/index.htm*.
7. Online tutorial of Ethernet, *http://wwwhost.ots.utexas.edu/ethernet/ethernet-home.html*.
8. Ethernet for the first mile (IEEE802.3ah Task Force), *http://grouper.ieee.org/groups/802/3/efm/index.html*.

BIOGRAPHY

Cedric F. Lam obtained his B.Eng. in Electrical and Electronic Engineering with First Class Honors from the University of Hong Kong in 1993. He finished his Ph.D. degree in Electrical Engineering from the University of California, Los Angeles (UCLA) in 1999 and joined AT&T Labs — Research as senior technical staff member of the Broadband Access Research Department. He has worked on a range of research projects, including fiber to the home (FTTH), hybrid fiber coaxial (HFC) systems, optical regional access networks, and optical signal modulation techniques. More recently, he has devoted his energy to the development and application of high-speed Ethernet technology in optical networking. In 2002, Dr. Lam joined Opvista Inc., where he is now project leader.

Dr. Lam received the AT&T Research Excellence Award for his contribution to the Metro-DWDM project in 2000. He was a recipient of the Sir Edward Youde Fellowship from 1994 to 1997 and a recipient of the UCLA Non-Resident Fellowship from 1995 to 1999. Dr. Lam is technical program chair of the 2002 Wireless and Optical Communication Conference (WOCC 2002), program committee chair of the 2002 Asian Pacific Optical and Wireless Communication Conference (APOC 2002) and Associate Editor of the OSA *Journal of Optical Networking*. He is a senior member of the Institute of Electrical and Electronics Engineers (IEEE).

BIBLIOGRAPHY

1. R. M. Metcalfe and D. R. Boggs, Ethernet: Distributed packet switching for local computer networks, *Commun. ACM* **19**(7): 395–404 (July 1976).
2. IEEE Standard 802.3, *Carrier Sense Multiple Access with Collision Detection* (*CSMA/CD*) *Access Method and Physical Layer Specifications*, 2000 edition.
3. A. S. Tanenbaum, *Computer Networks*, 3rd ed., Prentice-Hall, 1996.
4. J. J. Roese, *Switched LANs*, McGraw-Hill, 1999.
5. IEEE Standard 802.1D, *Information Technology — Telecommunications and Information Exchange Between Systems — Local Area Networks — Media Access Control* (*MAC*) *Bridges*, 1993.
6. IEEE Draft P802.3ae/D4.0, Dec. 2002.
7. U. Black and S. Waters, *SONET & T1: Architectures for Digital Transport Networks*, Prentice-Hall, 1997.
8. P. E. Green, Jr., *Fiber Optic Networks*, Prentice-Hall, 1993.
9. W. Xin, G. K. Chang, and T. T. Gibbons, Transport of Gigabit Ethernet directly over WDM for 1062 km in the MONET Washington DC network, *2000 Digest of IEEE LEOS Summer Topical Meeting on Broadband Optical Networks*, Aventura, FL, July 2000, pp. 9–10.
10. ITU-T Recommendations G.983.1, G.983.2, and G.983.3.
11. W. J. Goralski, *Introduction to ATM Networking*, McGraw-Hill, 1995.

MULTIBEAM PHASED ARRAYS

RANDALL G. SEED
MIT Lincoln Laboratory
Lexington, Massachusetts

1. INTRODUCTION

Multibeam phased arrays are used in wireless communications to establish RF links between one site and multiple other sites. The links may be either unidirectional or bidirectional. Multiple beams are formed from a phased-array antenna to provide directional point-to-point communications between nodes. The antenna beams from a phased array do not need to point in a constant fixed direction, as with a standard array antenna, or a reflector antenna. This property suggests that phased arrays are suitable not only for serving multiple users, but also for serving moving communications nodes. Multibeam phased arrays, in the current context, includes fixed and moving phased-array antenna beams (see Fig.1).

A phased-array antenna is composed of a large or small number of individual radiating antenna elements arranged in a one-dimensional (1D) or two-dimensional (2D) array. The 2D array is of more practical interest since it possesses the higher degree of freedom enabling greater beam agility. The 1D case is useful for mathematically demonstrating phased-array operation, and for illustration. Although generally the array is a flat plane, the elements can be arranged on a curved surface and thus made to conform to the shape of the vehicle that may host the antenna. Usually, all of the array's radiating elements are combined together to form one or more composite antenna beams from the array, and that will be assumed. This means that the multiple beams, producing the multiple channels, each emanate from the same single array, and are superimposed on each of the array elements. The signals from the elements are combined,

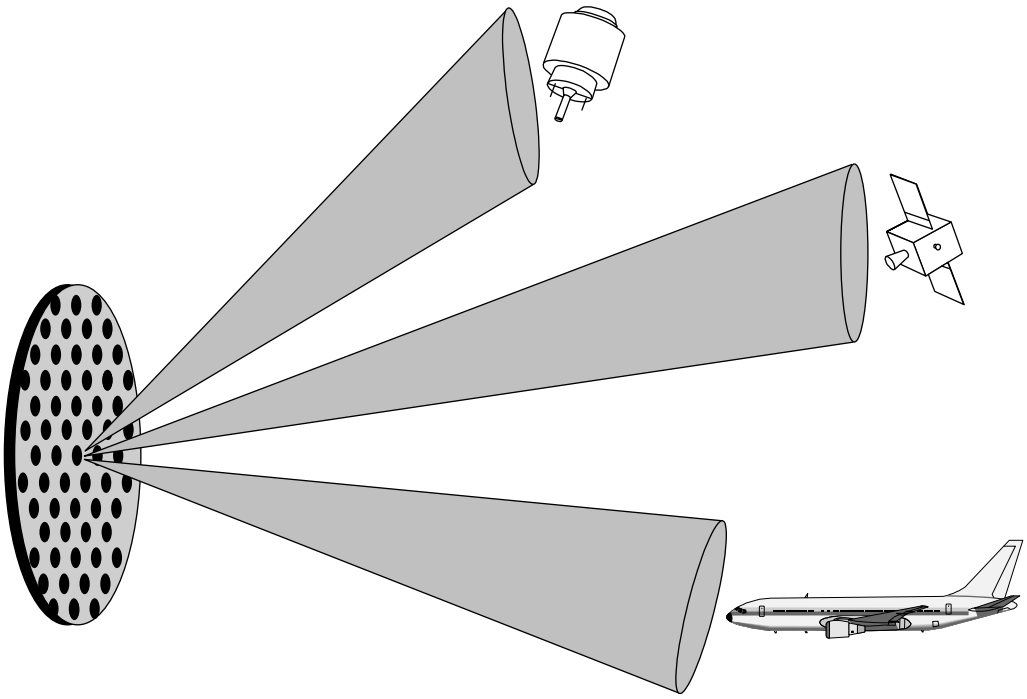


Figure 1. Single multibeam phased array antenna for simultaneous communications with multiple nodes.

or divided, depending on the direction of the signal, in a beam forming network. Each element can, and often does, radiate the distributed common RF signal with a different phase shift. The phase shift is defined relative to a reference element on the array. For convenience, the reference element is the nearest neighbor in a given direction. When the phase of the RF signal is altered on the individual radiating elements, the combined emanations form one or more directional beams in space. The beams are computed and measured in the "far field," also known as the *Fraunhofer region*. Nearer the array, in the "near field" or Fresnel region, the composite electromagnetic field structure is more complicated and generally more difficult to evaluate. The near-field analysis is not of relevance for communications link analysis, but rather for the antenna and antenna structural design.

A multibeam phased array antenna would generally be of interest in the following cases:

- Multiple users of undetermined angular position
- Broad area coverage with sustained high gain
- Multipath or jammer interference cancellation
- Multiple users, with node antenna volume and weight constraint
- Spectrum reuse

2. MULTIPLE-BEAM FORMATION

2.1. Single-Beam Linear Array

For a linear array, depicted in Fig. 2, the single beam array factor is

$$F(\theta) = \sum_{n=0}^{N-1} A_n e^{jkdn(\sin(\theta)-\sin(\theta_0))} = \sum_{n=0}^{N-1} A_n e^{jn(kd\sin(\theta)-\alpha)} \qquad (1)$$

when $A = 1$, then

$$|F(\theta)| = \frac{\sin\left[\frac{N}{2}kd(\sin(\theta)-\sin(\theta_0))\right]}{\sin[kd(\sin(\theta)-\sin(\theta_0))]} \qquad (2)$$

Example 1. Single beam from an unweighted linear array:

Array element spacing:	$d = \lambda/2$ m
Frequency:	$f = 1 \times 10^9$ Hz
Wavelength:	$\lambda = c/1 \times 10^9$ m
Number of elements:	$N = 16$ elements
Beam scan angle:	$\theta_0 = 7.18^\circ$ ($\alpha = 360^\circ/N$)
c is the speed of light	

The results are illustrated in Fig. 3.

Each 360° of total phase shift across the array, rotates the beam by one beamwidth. In this example, 7.18° is approximately one beamwidth.

2.2. Multibeam Linear Array

For a linear array producing two beams at the same frequency, the array factor is

$$F(\theta) = \sum_{n=0}^{N-1} A_{1n} e^{jkdn(\sin(\theta)-\sin(\theta_{10}))} + A_{2n} e^{jkdn(\sin(\theta)-\sin(\theta_{20}))} \qquad (3)$$

Since the two beams are assumed continuous wave, and are at the same frequency, the phases for each beam are additive and the signals to each beam position are correlated. As a result of this correlation, the beams will interfere unless suitable angular spatial isolation is provided. Frequency reuse may be obtained for modest

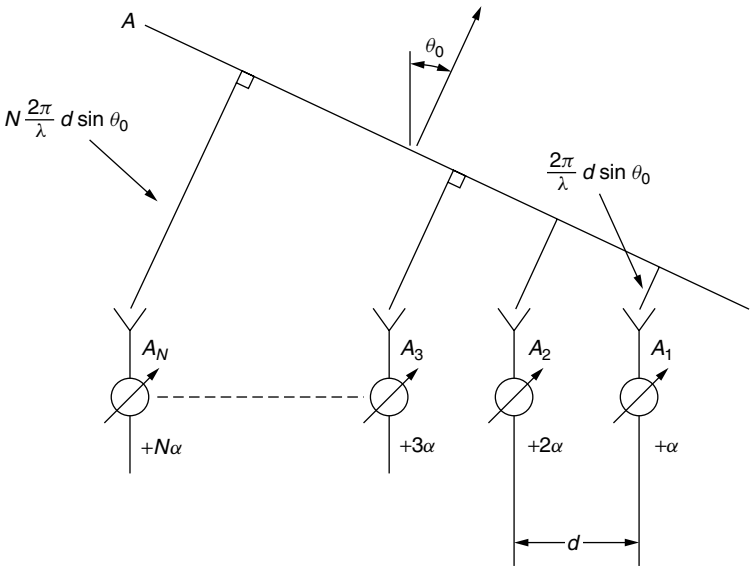


Figure 2. Geometry for linear array.

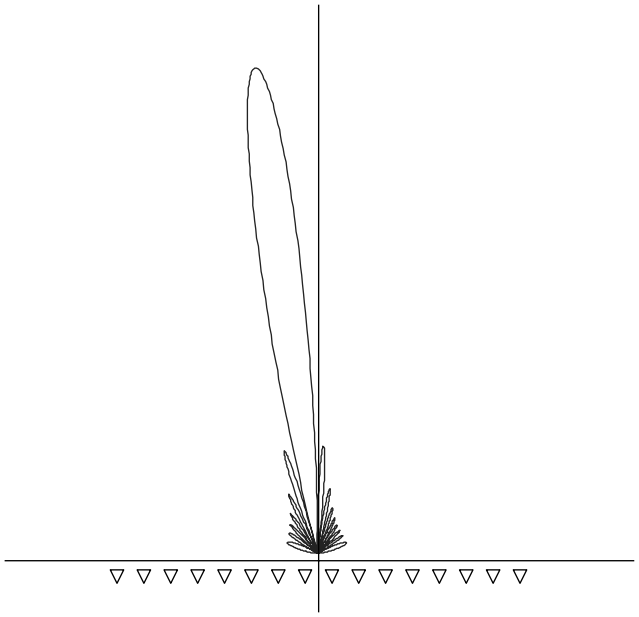


Figure 3. Single beam from a linear antenna array.

angle separation by providing suitable amplitude taper for each element, based on the amplitude taper for each beam, A_{1n}, and A_{2n}, such that the sidelobes from one beam negligibly impacts the other. Numerous amplitude weighting tapers are described in Chapter 2 of Hansen [1] for linear arrays, and Chapter 3 for planar arrays [1]. Spatially well separated beams will enable spectrum reuse. Finally, in the absence of an amplitude weighting taper, or spatial separation, then uncorrelated, or nearly uncorrelated signals will minimize the interference for arbitrarily spaced beams at the same frequency. Under this condition, frequency reuse channelization may be accomplished by generating uncorrelated signals within the same band, by means of polarization isolation, time-division multiplexing, or code-division multiplexing.

Example 2. Multiple beam, common frequency linear array.

$$d = \lambda/2$$
$$N = 16$$
$$\theta_{1,0} = 7.18^\circ \ (\alpha = 360/N)$$
$$\theta_{2,0} = -14.36^\circ \ (\alpha = -2 * 360/N)$$

In this example, two beams are generated and are separated by 3 beamwidths; the resulting pattern is illustrated in Fig. 4.

In design, it is best if signals are kept orthogonal, or minimally correlated for each beam, in order that spatial beams are separable, as in Fig. 5. In this case, the resulting beams are characterized by the individual array factors overlaid upon one another as in the graphic. Sidelobe levels are higher in Fig. 4 than in Fig. 5 because of phase interference.

2.3. Multibeam Two-Dimensional Array

For a two dimensional, planar rectangular array, the array factor is defined by

$$F(\theta) = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} A_{mn} e^{\begin{array}{l} jk[md_x(\sin(\theta)\cos(\phi)-\sin(\theta_0)\cos(\phi_0)) \\ +nd_y(\sin(\theta)\sin(\phi)-\sin(\theta_0)\sin(\phi_0))] \end{array}}$$
$$= \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} A_{mn} e^{jk[md_x(u-u_0)+nd_y(v-v_0)]} \qquad (4)$$

For uncorrelated signals, the antenna beams are non-interfering, and the multiple beam patterns are shown as graphical overlays of each array factor antenna pattern.

Example 3. Multibeam generation from planar rectangular array:

$$d_x = d_y = \lambda/2$$
$$N = M = 16$$

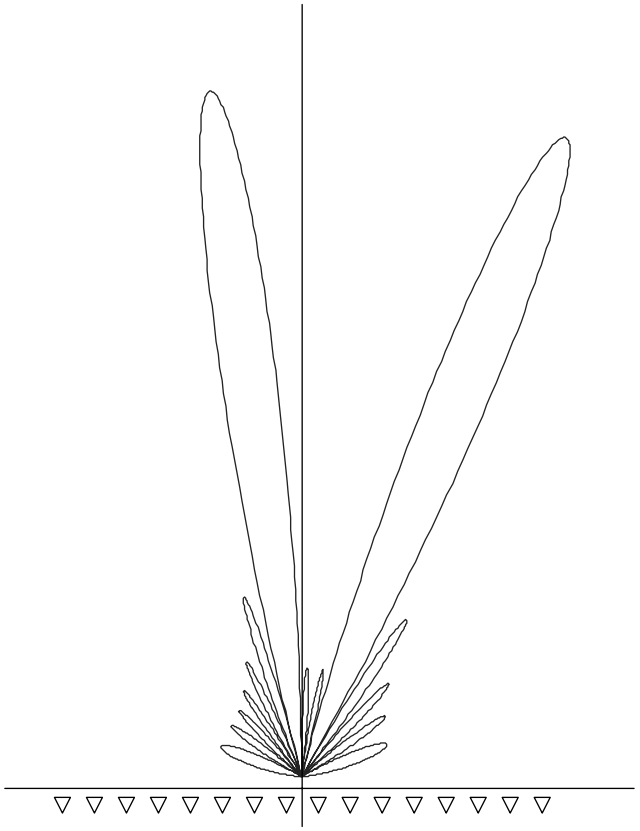


Figure 4. Multiple correlated beams from a single 16-element linear antenna array.

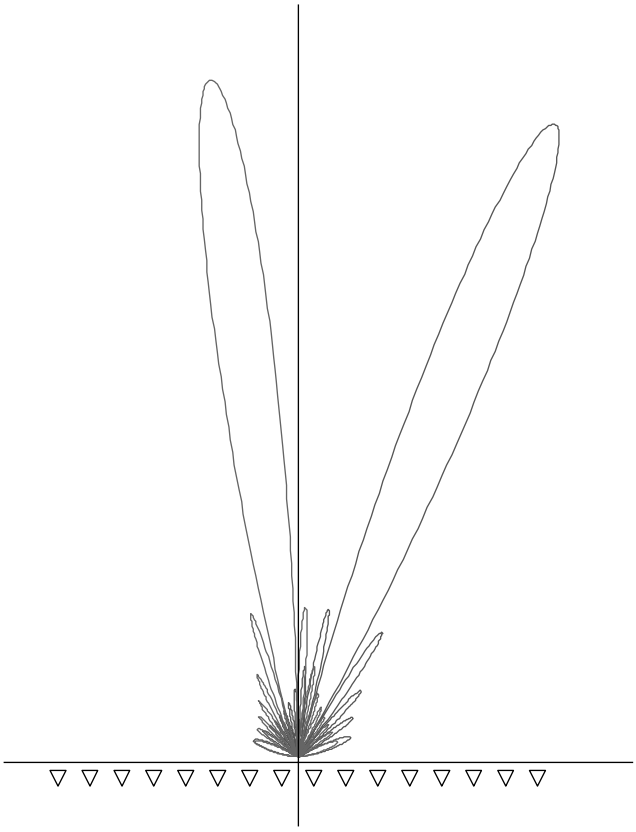


Figure 5. Multiple, uncorrelated beams, generated from a single linear antenna array.

$\theta_{1,0} = 0^\circ$
$\phi_{1,0} = 0^\circ$
$\theta_{2,0} = 20^\circ$
$\phi_{2,0} = 45^\circ$
$f = 1 \times 10^9$ Hz
$\lambda = c/1 \times 10^9$ m

A 3D representation of the resulting dual-beam pattern from the rectangular 2D array is plotted in u, v coordinates in Fig. 6.

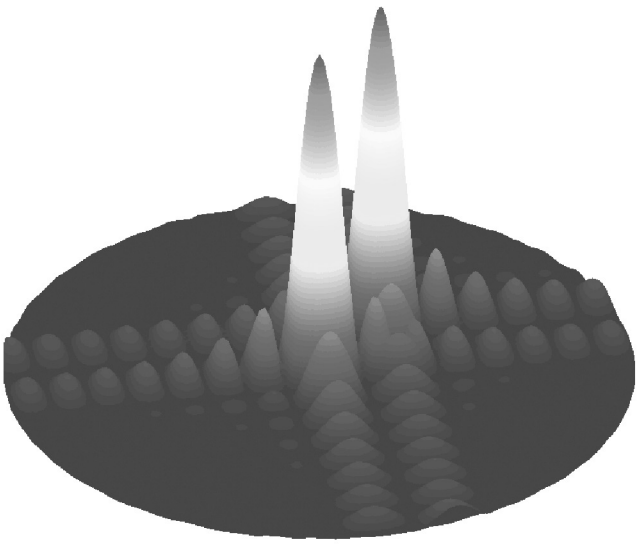


Figure 6. Multiple uncorrelated beams from a single two-dimensional phased array plotted in u, v coordinates.

There is a continuum of beam positions that can be obtained from the phased-array system. However, in practical terms the maximum number of beams required is the number of beams that fill the assigned angular volume, given an acceptable fractional beam overlap. The typical overlap is chosen as the angular distance at which adjacent beams are 3 dB below their peak gain value.

Tolerable antenna beam scan loss, and mutual coupling of antenna elements generally limit the designer's choice to scanning coverage of plus or minus 60°. Then, since the 3-dB beamwidth is

$$\theta_{3\ \mathrm{dB}} \cong \frac{0.886\lambda}{L} \tag{5}$$

for a linear array with uniform amplitude, the number of beams required to fill the angular space with 3-dB overlap per beam is

$$n \sin\left(\frac{0.886\lambda}{L}\right) = 2\sin(60^\circ) \tag{6}$$

For half-wavelength element spacing, we find that $n = N - 1$. By forcing the beam peaks for the end beams to

coincide with the edge of coverage, then the number of beams that fill the volume is equal to the number of elements, $n = N$.

Now, given this information, the choice to be made for beamforming is from:

1. Producing the logic and electronics to compute the phase shifts on each element to generate the desired number, M, of arbitrary beams
2. Hard-wiring, or hard-coding, the phase shifts per each element to form N beams from which the user selects M

A general beamformer used to produce N beams from an N element array is shown in Fig. 7. The number of electronic components used to form the beams is N attenuators and N phase shifters for each beam. This network requires $N \times N$ signal combinations, for a total of N^2 signal combinations.

On the other hand, with the second option for fixed beamforming, there are many applications in which it is convenient to form the fixed beams, and choose from among these N beams. The Butler matrix, for example, is used to form N beams using a minimum of phase shift and signal combining elements. The Butler beamforming matrix has been described as analogous to the FFT (fast Fourier transform) when N is a power of 2, and is as shown in Fig. 8 [1]. It requires only $N \times \log(N)$ combinations to form N beams. All lines are equal length, except for ones with phase shift.

2.4. Design Considerations for Multibeam Phased Arrays

The major design considerations for multibeam phased arrays for communications, are as follows:

Gain. Driven by the signal to noise ratio required by the two ends of the link. Drives array size, beamforming methodology.

Beamwidth. Driven by the requirement for spatial isolation of beams. Drives array size, frequency.

Scan Loss. Losses at maximum scan angle from array face normal. Drives array size.

Sidelobe Level. Driven by requirement for spatial isolation and interference rejection. Drives beamforming method, amplitude taper requirement.

Number of Simultaneous Users — Transmit. Driven by link requirements. Drives beamforming method, output power of array elements.

Number of Simultaneous Users — Receive. Driven by link requirements. Drives beamforming method.

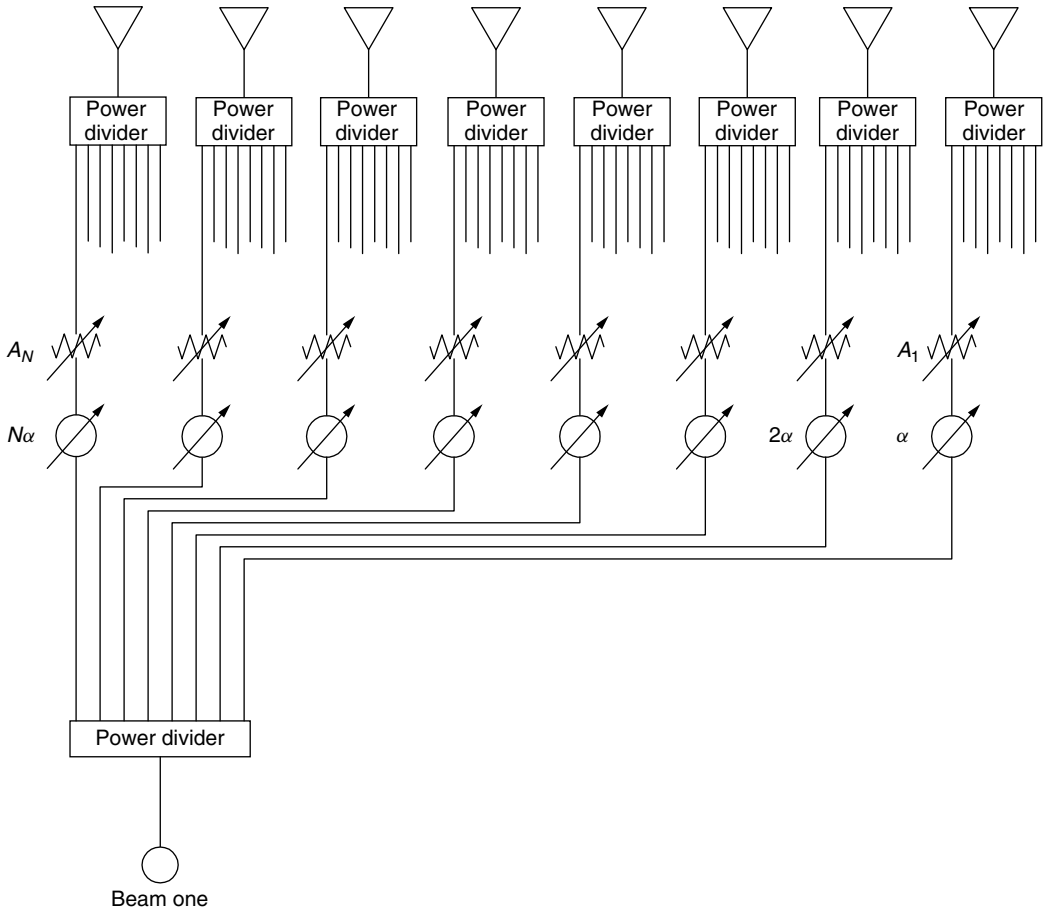


Figure 7. General beamformer for an N-element array.

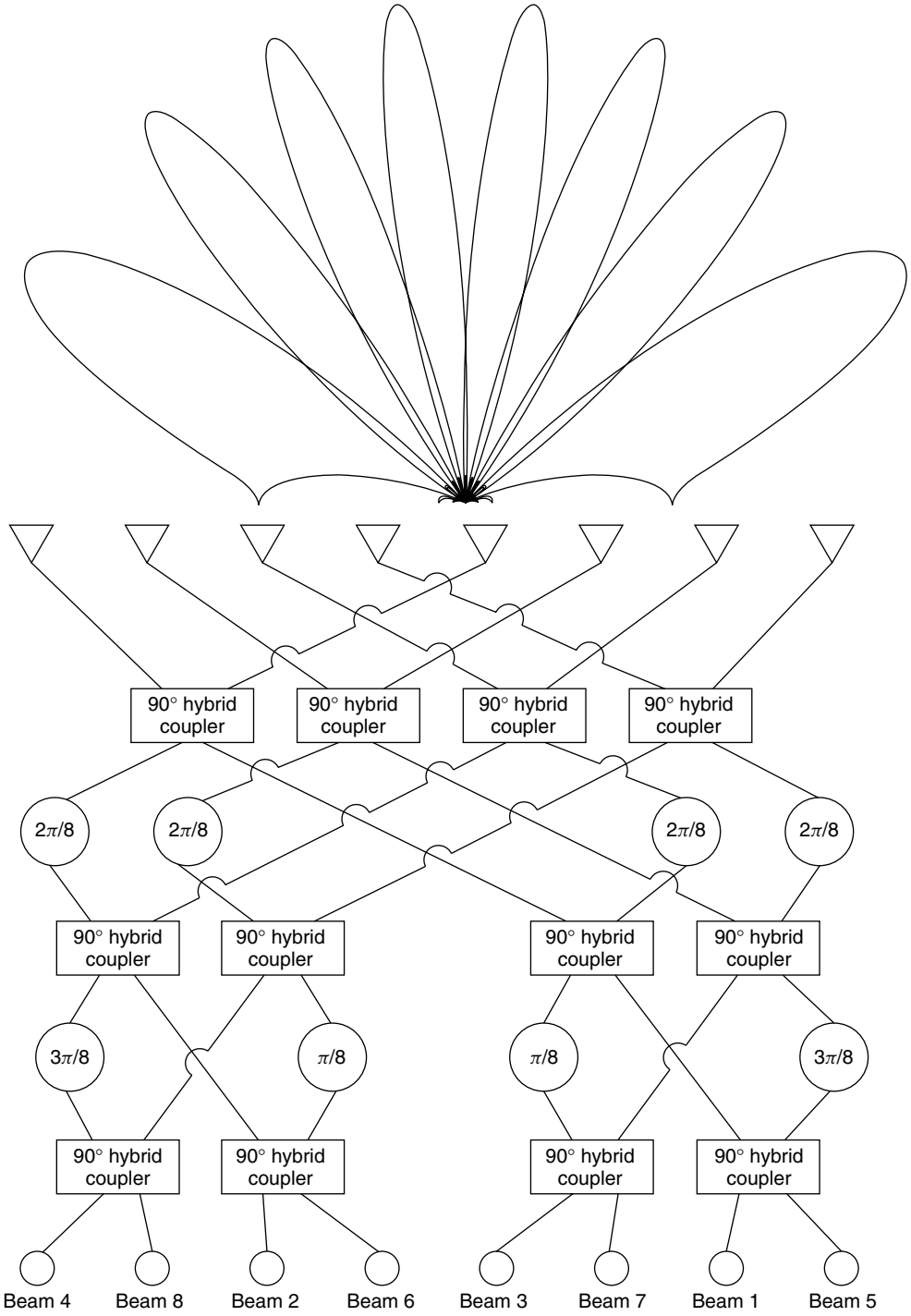


Figure 8. Butler beamformer for an 8-element linear array antenna.

Bandwidth. Driven by link requirement. Drives gain, scan loss, beamwidth, sidelobe level, and, beamforming method.

Generally, cost is the major issue with the phased-array antenna from an overall design perspective, versus a reflector antenna or fixed array. This is due to the number of amplifiers, transmit/receive switches, splitters, combiners, and the whole of the beamforming network. From an engineering point of view, the most significant issue is bandwidth. Since the array is driven by phase shifters, and not true time delay elements, a transmitted signal will undergo dispersion across the array face. The dispersion becomes a factor, and increases as the beam is scanned off of normal (see PHASED-ARRAY ANTENNAS).

Sensitivity on receive is set by the noise figure of the first-stage array element amplifier. The signal from an array element can be split as many times as desired after

the first stage without effective degradation in the SNR. This suggests the option of a large number of signals and channels with which to process the antenna beams. If the data are digitally sampled after the first stage, the signals can be combined in a computer.

Maximum transmit capability is defined by the maximum power output of the array element module. Since the signals are additive in the final-stage amplifier, fixed input levels are imposed. An option to take advantage of the full gain of the final-stage amplifier, is to time-division-multiplex the channels during transmission from the phased array. Both signal limiting and multiplexing of transmission are employed in practice.

3. EXAMPLES AND CURRENT APPLICATIONS

It is evident from the discussion that one advantage gained in developing a multibeam phased-array antenna is that one physical phased-array antenna with M beams, nearly performs the same task as M independent antennas. This advantage is appreciated, and exploited in the development of earth orbiting spacecraft, where size and weight become dominant constraints over complexity. The first communications systems to use multibeam phased arrays included several space telecommunications programs. Two of these are summarized, one U.S. government program, and one commercial program.

3.1. Government — TDRS System

The Tracking and Data Relay Satellite (TDRS) program, run by NASA, was an early user of the multiple beam phased array for communications. It is probably the best example available today, and certainly of its time, of multiple adaptive phased-array antenna beams used for communications.

The TDRS system is a series of spacecraft, operating independently from one another, that are positioned in geosynchronous orbit. Each spacecraft's set of antennas includes a phased array that points continuously towards the earth. It is designed to be able to acquire and communicate with other orbiting satellites or space vehicles, primarily those at low to medium earth orbit. The array has a field of view of plus or minus 13.5°, which allows it to follow satellites at up to 2300 mi altitude when not earth-eclipsed.

The TDRS multiple-access (MA) phased-array antenna is composed of 30 helical antenna array elements that, when combined, produce beams of nominal size $3.6^\circ \times 2.8^\circ$. All users transmit at the same frequency in 6 MHz bandwidth using QPSK (quadrature phase shift keying) at 2.2875 GHz. Their CDMA code key distinguishes users. Each TDRS user has a unique pseudonoise Gold code, yielding signals that are nearly uncorrelated. This minimizes antenna beam interference for unpredictable user position, and allows for the separation of individual user messages in the code domain [2].

The TDRS system avoids the need for complex beamforming circuitry in space, by transmitting the full bandwidth of each of the 30 elements to the TDRS ground terminal. Each element's 6 MHz information band is multiplexed into one of thirty 7.5-MHz channels and relayed to the ground in 225 MHz bandwidth. The ground system demultiplexes and digitally samples each channel. The digital antenna array element information is combined and reconstructed to form antenna receive beams in the direction of selected user satellites [3].

The TDRS MA beamforming process is executed in three modes:

1. *Beam steering* by knowing the position of the user beforehand, and setting the element phase shifts appropriately
2. *Adaptive beam steering*, determining the angles from measurements and signal processing
3. *Interference canceling*

One possible advantage to the adaptive method, is that a satellite's angular position may be determined by adaptively processing the returns employing the user's PN key, and thus, the vehicle may be closed-loop tracked.

In theory, any number of beams can be formed from the downlink data, since the raw measurements are available for each element. Beam formation limitations arise in the processing electronics at the ground station and, in this case, are reportedly limited to 10 beams per each of two ground terminals, for a total of 20 beams. Also, the number of elements, N, in this case 30, limits the ability to cancel co-channel interference. The array possesses $N - 1$ degrees of freedom [see Adaptive Arrays]. There are nominally four satellites in the constellation and so 5 beams are allocated to each spacecraft.

The second generation advanced TDRS system is under development now. The first and second of three satellites were launched in 2000 and 2001. Among other enhancements, the new TDRS system has moved the receive beamforming function of the multiple access phased-array antenna, onto the spacecraft itself [5].

3.2. Commercial — Iridium

The Iridium system is a constellation of 66 low-earth-orbit (LEO) satellites (the original system constellation was designed to contain 77 — the atomic number of iridium).

Each Iridium satellite has three main mission antenna (MMA) panels, each with approximately 108 patch antenna elements in a two-dimensional array. Sixteen simultaneous fixed beams are formed in two steps in the beamforming unit. First, 80 beamlets are formed using two-dimensional crossed Butler beamforming matrices. The Butler beamformer is composed of eight 16×16 Butler matrices, followed by ten 8×8 Butler matrices. Secondly, the sixteen beams are formed by performing power division/combining on groups of the 80 beamlets [6].

The communications channels are broken into 120 FDMA channels. Transmit and receive operations are effected by means of transmit/receive modules and separated using TDMA with four transmit and four receive channels for each FDMA channel [7]. Beam positions are switch selected from among the outputs of the fixed beamforming network.

4. DIGITAL BEAMFORMING

Multibeam phased array development trends are toward advanced techniques using digital beamforming, which enables adaptive beam steering, and adaptive nulling. This technique can be represented by a simple change to the block diagram: the amplitude and phase weights are performed digitally, therefore, the signal is digitally sampled at each element.

The technology driver for digital beamforming lies in the A/D (analog/digital) electronics. Although dynamic range requirements on digital sampling are less restrictive when conversion is performed at the element level, the A/D converter must operate at a rate of twice the full bandwidth of the entire link. For broadband, frequency-division multiple access communications, these rates can be formidable. Additionally, the datastream includes the samples from all of the N elements at once, and therefore, the internal data rate before signal processing becomes N times the sampling rate times the number of bits per sample.

Major advantages of digital beamforming include (1) full beamshape control, (2) ability to produce true time delay (vs. phase shift), by means of digital delay, (3) interference cancellation, and (4) adaptable calibration.

4.1. Adaptive Beam Steering

Adaptive beam steering is performed by using the least-mean-squares (LMS) beam-steering algorithm [8]. Beam steering can be performed for each of a multiple number of signals in a multiuser system. A signal, s, impinges the array face, producing the response, x, on each of 1 to N elements:

$$\mathbf{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \tag{7}$$

Compute the signal-plus-noise covariance matrix for the excitation on the array elements, M:

$$\mathbf{M} = E[\mathbf{X}^*\mathbf{X}^T] \tag{8}$$

Compute a reference vector for the signal of interest as

$$\tilde{\mathbf{S}} = E[\mathbf{X}^* r] \tag{9}$$

The reference signal contains the representation, $r(t)$, of the desired signal, for example, that representation may contain the PN code for the channel of interest.

Now, the weight vector, W, containing the appropriate antenna element phase shifts to steer the beam in order to maximize the signal, is defined by

$$\mathbf{W} = \frac{\mathbf{M}^{-1}\tilde{\mathbf{S}}}{\tilde{\mathbf{S}}^{*T}\mathbf{M}^{-1}\tilde{\mathbf{S}}} \tag{10}$$

including a normalizing factor in the denominator and the best estimate signal is

$$\hat{s} = \mathbf{W}^T\mathbf{X} \tag{11}$$

The elements, W, form the array phase shifts, and amplitude weights, for the estimated beam.

Example 4. Single beam, linear array, adaptively steered. Let $N = 16$. Then

Signal at $\theta_{1,0} = 9.6^\circ$ ($\alpha = \pi/6$)
Signal power at each element = 1
Noise power at each element = 1
SNR at each element = 1.0 → 0 dB (see Fig. 9)

4.2. Adaptive Interference Cancellation

The Applebaum array is used for reducing, or canceling, the effects of interference, [4,8]. The procedure is analogous to adaptive beam steering, however, the covariance is computed slightly differently, as is the steering vector, **S**.

The noise covariance matrix is computed for the element response:

$$\mathbf{M}_n = E[\mathbf{X}_n^*\mathbf{X}_n^T] \tag{12}$$

Set the steering vector, **S**, to the phases of the elements for the current beam pointing position:

$$\mathbf{S} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_N \end{bmatrix} \tag{13}$$

Figure 9. One statistical representation of single beam adaptive beamforming of a 16-element linear array with 0 dB SNR per element. Adaptively formed beam — blue; ideal beam — red.

If it is a linear array, we obtain

$$\mathbf{S} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_N \end{bmatrix} = \begin{bmatrix} e^{-j\alpha} \\ e^{-j2\alpha} \\ \vdots \\ e^{-jN\alpha} \end{bmatrix} \tag{14}$$

Now, the weight vector, W, containing the appropriate antenna element phase shifts to modify the beam in order to maximize the signal in the presence of interference, is defined by

$$\mathbf{W} = \mathbf{M}_n^{-1}\mathbf{S}^* \tag{15}$$

and the maximized signal is

$$\hat{s} = \mathbf{W}^T\mathbf{X} \tag{16}$$

[see also Adaptive Arrays].

5. MULTIBEAM PHASED ARRAYS — FUTURE

Current trends in multibeam phased arrays indicate that future development is aligned toward digital beamforming. Key areas of focus include high-speed analog-to-digital converters. Impetus originates primarily from military users, including digital beamforming for radar and for multiuser spread-spectrum communications. The commercial world has its primary motivation for development of multibeam phased arrays for continued applications in space.

BIOGRAPHY

Randall Graham Seed received the B.S., M.S., Ph.D. degrees in electrical engineering from Northeastern University, Boston Massachusetts in 1990, 1992, and 1994, respectively. He was a Visiting Assistant Professor at Northeastern University from 1994 to 1995. In 1995 he was employed by TRW, Redondo Beach, California, where he worked on research and development in electromagnetic applications and space systems engineering. He joined Raytheon, Bedford, Massachusetts, in 1998, performing systems engineering on strategic ground-based radars. Since 2001, he has been with MIT Lincoln Laboratory, Lexington, Massachusetts, where he is engaged in research, design, analysis and development of novel airborne, sea-based, ground-based, and space based RF sensor systems. He has authored or co-authored approximately 20 papers, primarily in RF materials technology. Dr. Seed is a licensed Professional Engineer in the state of California, and is a member of IEEE. His areas of interest include advanced phased-array technology for radar and communication systems, sensor systems engineering, and active and passive materials technology applications to RF systems.

BIBLIOGRAPHY

1. R. C. Hansen, *Phased Array Antennas*, Wiley, New York, 1997.
2. R. Avant, B. Younes, D. Lai, and W.-C. Peng, STGT multiple access beamforming system modelling and analysis, *Military Communications Conf., 1992, IEEE MILCOM '92, Conf. Record, Communications — Fusing Command, Control and Intelligence*, 1992, Vol. 3, pp. 1028–1034.
3. R. Avant, B. Younes, G. Dunko, and S. Zimmerman, STGT multiple access beamforming equipment PC analysis system, *Military Communications Conf., 1992, IEEE MILCOM '92, Conf. Record, Communications — Fusing Command, Control and Intelligence*, 1992, Vol. 3, pp. 1035–1039.
4. S. P. Applebaum, Adaptive arrays, *IEEE Trans. Antennas Propag.* **24**(5): 585–598 (1976).
5. Space Network Online Information Center (no date), NASA Goddard Space Flight Center, *http://nmsp.gsfc.nasa.gov/tdrss/* (Jan. 2002).
6. J. J. Schuss et al., The IRIDIUM main mission antenna concept, *IEEE Trans. Antennas Propag.* **47**(3): 416–424 (1999).
7. R. A. Nelson (no date), Iridium: From Concept to Reality (online), Applied Technology Institute, *http://www.aticourses.com/news/iridium.htm* (Jan. 2002).
8. R. T. Compton, Jr., *Adaptive Antennas*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

FURTHER READING

Brookner E., ed., *Practical Phased Array Antenna Systems*, Lex Book, Lexington, MA, 1997.

Compton R. T., Jr., An adaptive array in a spread-spectrum communication system, *Proc. IEEE* **66**(3): 289–298 (1978).

Hansen R. C., ed., *Microwave Scanning Antennas*, Peninsula Publishing, Los Altos, CA, 1985.

Iridium Home, (2001). [Online]. Iridium Satellite LLC, *http://www.iridium.com/* (Jan. 2002).

Mailloux R. J., *Phased Array Antenna Handbook*, Artech House, Boston, 1994.

Zaghloul A. I., Y. Hwang, R. M. Sorbello, and F. T. Assal, Advances in multibeam communications satellite antennas, *Proc. IEEE* **78**(7): 1214–1232 (1990).

MULTICARRIER CDMA

DIMITRIS N. KALOFONOS
Northeastern University
Boston, Massachusetts

1. INTRODUCTION

The term *multicarrier CDMA* (MCCDMA) is used to describe multiple-access schemes that combine multicarrier modulation (MCM) and code-division multiple access (CDMA) based on direct-sequence spread spectrum (DSSS). Different ways of combining MCM and DSCDMA were proposed in 1993 by a number of researchers independently, and since then this idea has attracted significant attention because it combines two very successful techniques. Most of the proposed MCCDMA schemes use orthogonal carriers in overlapping subchannels for multicarrier transmission, also referred to as *orthogonal frequency-division multiplexing* (OFDM), but some MCCDMA schemes use few nonoverlapping subchannels. A MCCDMA scheme using nonoverlapping subchannels

is less bandwidth-efficient, but its implementation is a straightforward extension of existing DSCDMA systems and backward-compatible with them. For these reasons this type of MCCDMA has been selected as one of the options for digital cellular third-generation (3G) CDMA systems. On the other hand, MCCDMA schemes using OFDM are more bandwidth-efficient because they allow for minimum carrier separation and can be efficiently implemented in DSP using the fast Fourier transform (FFT). For these reasons OFDM-based MCCDMA schemes have attracted more intense research interest, although their adoption in practical systems is still limited.

MCCDMA schemes share many of the characteristics of both MCM and DSCDMA and attempt to make use of features of one component to overcome limitations of the other. The first component of MCCDMA, which introduces frequency-domain spreading in the signal design, is MCM [1]. MCM is based on the principle of transmitting data by dividing the stream into several parallel bitstreams, each of which has a much lower rate, and using these substreams to modulate several carriers. In this way, the available bandwidth is usually divided into a large number of subchannels, and each substream is transmitted in one of these subchannels. Since MCM is a form of frequency-division multiplexing (FDM), earlier MCM design borrowed from conventional FDM technology and used filters to completely separate the subchannels. This approach was soon abandoned since very sharp cutoff filters were needed and the number of subchannels that could be implemented was very small. The breakthrough in implementation of MCM came when the spectra of the individual subchannels were allowed to overlap, but the signals could still be separated at the receiver because they were mutually orthogonal. It was also found that in this approach both transmitter and receiver can be implemented using efficient fast Fourier transform (FFT) techniques. The orthogonality of the signals transmitted in different subchannels is achieved by using carrier separation of

$$\Delta f = \frac{1}{T_b} \tag{1}$$

where T_b is the MCM symbol duration. Thus, the carriers are located in frequencies

$$f_k = f_0 + k\Delta f, \quad k = 0, 1, \ldots, N-1 \tag{2}$$

where N is the total number of the subchannels and Δf is given in (1). Conceptually the transmitter and the receiver have the structure depicted in Fig. 1. In practice, MCM is implemented using the FFT, as depicted in Fig. 2, where block diagrams of the transmitter and the receiver are shown. With a proper selection of the guard interval, it can be shown that the effect of the channel on the transmitted symbols X_i is a multiplicative complex coefficient, which is the frequency response of the channel in the range of the respective subchannel. Note that by selecting the MCM symbol duration T_b large enough, the subchannels become narrow enough, so that the response of the channel in each of them can be considered approximately flat (non-frequency-selective channel)

$$Y_i = H_i X_i + N_i \tag{3}$$

where H_i is the complex channel coefficient for subchannel i, and N_i is the AWGN.

MCM enables high data rates and efficient bandwidth utilization and at the same time allows for large symbol duration. This large symbol duration is the most important characteristic of MCM, because it allows for almost intersymbol interference (ISI)-free transmission in both fixed and randomly fading frequency-selective channels. Other advantages of MCM are approximately flat-fading subchannels, which facilitate the inversion of the effects of the channel easier, and efficient and flexible implementation. On the other hand, the most serious drawbacks of MCM are sensitivity to carrier synchronization, nonlinear distortion because of the high peak-to-average values, and vulnerability to frequency selective fading, which can only be alleviated through signal diversity obtained with coding and interleaving in the frequency domain.

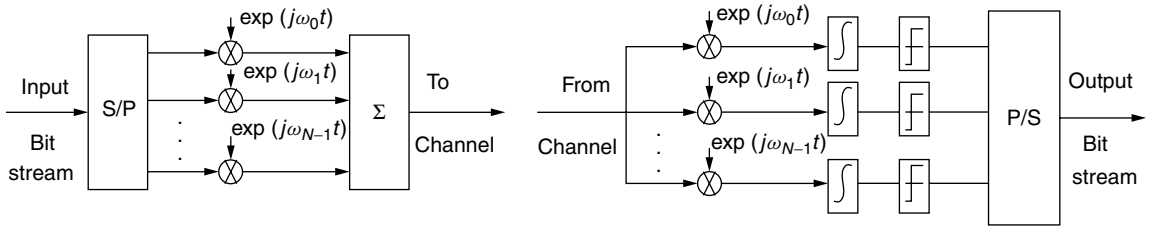


Figure 1. Conceptual structure of MCM transmitter (left) and receiver (right).

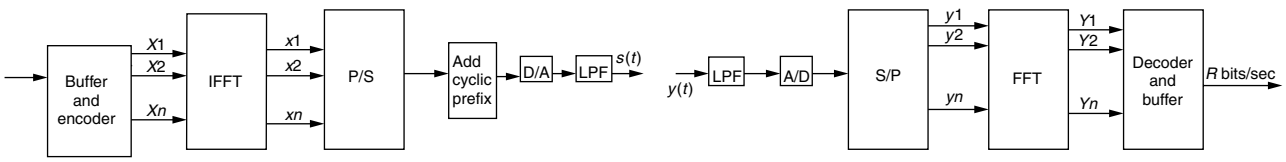


Figure 2. Implementation of MCM transmitter (left) and receiver (right) using FFT.

The other component of MCCDMA, which introduces time-domain spreading in the signal design, is DSSS [2]. DSSS signals are characterized by the fact that their bandwidth W is much greater than the information rate R. The redundancy added gives DSSS some important properties that make this technique popular for military and commercial applications. DSSS systems are used for combating interference due to jamming, other user interference, and self-interference caused by multipath propagation. The ability of DSSS systems to cope with interference caused by other users using the same channel makes DSSS the basis of a very effective multiple-access technique, namely, DSCDMA. In a DSCDMA system each user is coding its information symbols using a pseudorandom sequence, called a *spreading sequence*. Usually these sequences are selected such that they have an impulselike autocorrelation function and very low maximum cross-correlation values. In a DSCDMA system, each user i, $i = 1, \ldots, N_u$, spreads each information bit $b_i(k), k = 0, 1, \ldots$, by using a spreading sequence signal of length N_s, and the spread signals of all users are added when transmitted through the channel. If the channel is a frequency-selective, multipath fading channel, each station will receive multiple echoes of the transmitted signals with different attenuations arriving from N_p different paths. The optimal receiver of such a system involves a filter matched to the convolution of the channel impulse response with the transmitted signal. An approximation to that receiver is the RAKE receiver with N_p fingers. A simplified block diagram of a DS-CDMA system with a RAKE receiver, is depicted in Fig. 3.

DSSS-based CDMA is a very successful multiple-access technique, that has been selected as the basis of many contemporary wireless systems. The most important of its characteristics is its superb capability to resist interference: narrowband from intentional or unintentional jamming, wideband from other users using the same frequency band, and self-interference from transmission in dispersive multipath fading channels. DSSS also allows for hiding a signal from undesired listeners and achieving privacy. Cellular systems take advantage of the DSCDMA receiver structure to achieve soft handoff. On the negative side, practical DSCDMA receivers have a limited number of RAKE fingers and cannot exploit all useful signal energy, which may reduce their performance. Accurate synchronization and continuous tracking of signal arrivals from different paths is also necessary.

Depending on the system design, MCCDMA systems can share more characteristics with either of the two system components. Examples can range between the MCCDMA system selected as an option in some 3G CDMA cellular systems, which resembles more a conventional DS-CDMA system; and MCCDMA systems, where all the spreading takes place in the frequency domain, which resemble more a conventional MCM system. Different MCCDMA schemes attempt to a different degree to combine the interference rejection capabilities inherent in DSCDMA and the bandwidth efficiency and long symbol duration inherent in MCM.

2. MCCDMA SCHEMES

Because it combines MCM and DSSS, MCCDMA offers the unique possibility to spread the original data in both the time and the frequency domains. The different MCCDMA schemes that have been proposed in the literature cover the range between conventional DSCDMA, which offers maximum spreading in the time domain and no spreading in the frequency domain, and the scheme we will refer to as OFDM-CDMA, which offers no spreading in the time domain and maximum spreading in the frequency domain.

The first scheme we examine was proposed by DaSilva and Sousa [3]. In this scheme, the available bandwidth is divided into N subchannels that correspond to orthogonal carrier separation. Each user creates a block of $\mu = N$ symbols, and each of these symbols is spread using the user's spreading sequence. The chips corresponding to the spread symbol are then transmitted over one of the available subchannels using MCM. Note that in this way, each MCM block symbol contains one chip from each spread symbol, so that the transmission of each spread symbol is completed after N_s subsequent multicarrier blocks, where N_s is the length of the user's spreading sequence. The transmitter structure

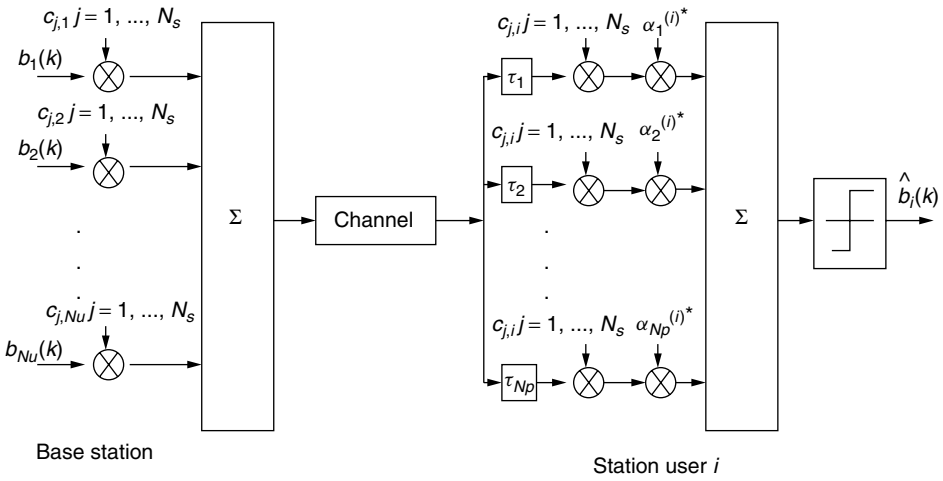


Figure 3. Block diagram of a DSCDMA system with a RAKE receiver.

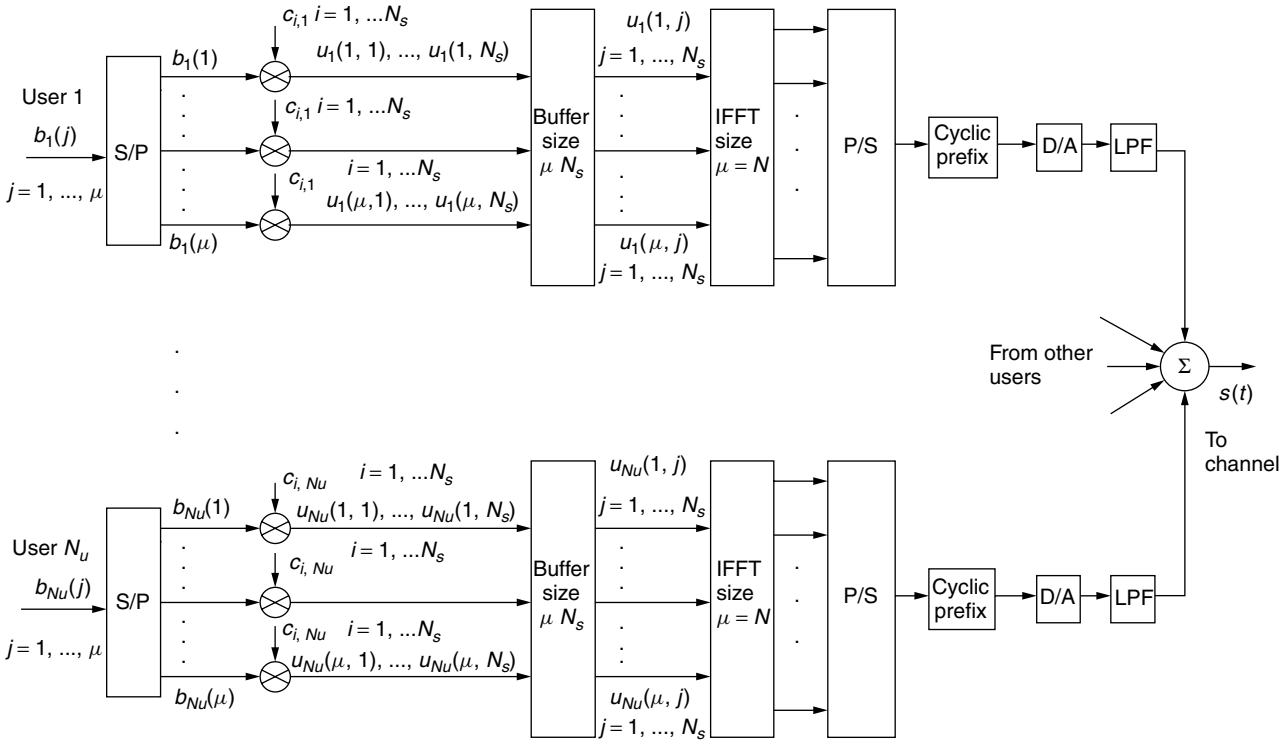


Figure 4. DaSilva and Sousa MCCDMA scheme: transmitter structure.

for this scheme is depicted in Fig. 4. This scheme uses multicarrier transmission as a way of maintaining the same data rate and spreading gain with a comparable DSCDMA system, while increasing the chip duration by a factor of N. This enables quasisynchronous transmission among different users and results in almost flat-fading subchannels, eliminating the need for RAKE receivers. The transmission of each symbol resembles that of a narrowband DSCDMA system over a flat-fading channel. Since each symbol is transmitted over only one subchannel, this scheme offers no frequency diversity. All the spreading is done in the time domain, and no spreading takes place in the frequency domain. Therefore, extensive coding of the symbols that are transmitted in parallel channels, often referred to in OFDM as *frequency-domain coding*, is needed to achieve acceptable performance [4]. Also, because of the large duration of the transmission of each symbol (N_s times the long MCM block symbol duration T_b), this scheme is more appropriate for slowly fading channels.

A variation of this scheme was proposed at the same time and independently by Kondo and Milstein [5]. The transmission of symbols in the available subchannels is done in a similar manner, but multiple copies of each symbol are transmitted in parallel in different subchannels as a means of introducing frequency diversity. The concept of the transmitter structure for this MCCDMA scheme is depicted in Fig. 5. In this scheme each symbol is spread in both the time and frequency domains. Although orthogonal carriers are considered in the original proposal, the authors later advocated the use of nonoverlapping subchannels as a practical design in realistic systems [6]. A scheme similar to that was later adopted as a multicarrier option for some 3G CDMA systems. A similar scheme that generalizes the idea of spreading symbols in both the time and frequency domains was proposed by Sourour and Nakagawa [7]. This scheme proposes the transmission of multiple copies of each symbol in orthogonal carriers and interleaving to maximize the achieved time and frequency diversity. This is a flexible system that allows the system designer to adjust the tradeoff between spreading in the frequency and time domains. Note that the term *multicarrier DSCDMA* (MCDSCDMA) is sometimes used [8] to distinguish these schemes, which allow for spreading in both the frequency and time domains, from other MCCDMA schemes.

The scheme proposed independently by Fazel [9], Chouly et al. [10], and Yee et al. [11] represents the other extreme in time–frequency spreading design. This scheme spreads each symbol entirely in the frequency domain, and there is no spreading in the time domain. The basic idea is to spread each symbol in the frequency domain by

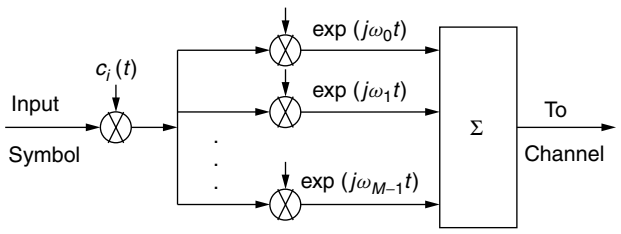


Figure 5. Kondo and Milstein MCCDMA scheme: transmitter structure.

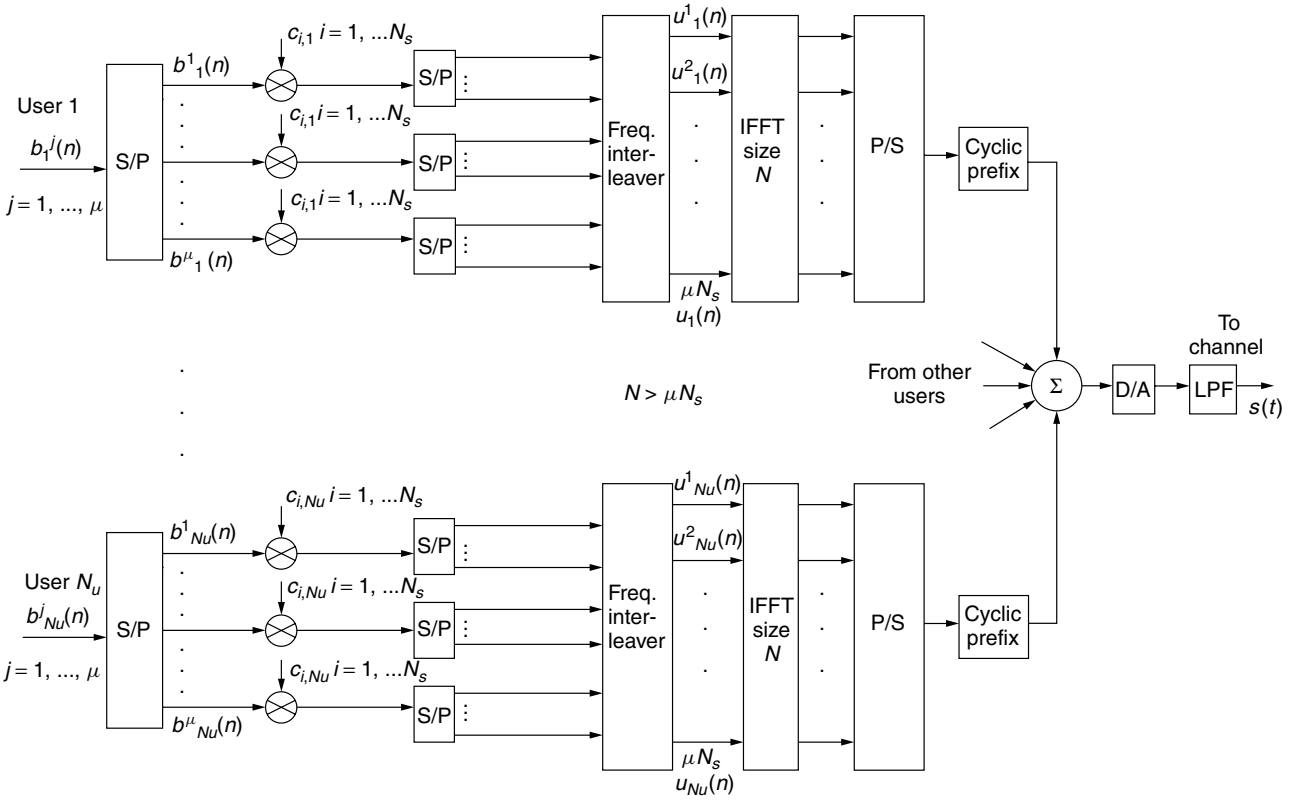


Figure 6. Fazel–Chouly–Yee [9–11] MCCDMA scheme: transmitter structure.

transmitting all the chips of a spread symbol at the same time, but in different orthogonal subchannels. The entire symbol is transmitted in one MCM block. This scheme offers the most frequency diversity among all MCCDMA schemes, since each symbol is transmitted in parallel over a large number of subchannels. To maximize the frequency diversity benefit, μ symbols of each user are transmitted in parallel in each MCM block symbol, where μ/T_b should be larger than the coherence bandwidth of the channel. Then with the addition of a frequency interleaver all chips corresponding to one data symbol are transmitted over subchannels undergoing approximately independent fading. The transmitter structure for this scheme is depicted in Fig. 6. This MCCDMA system has attracted the largest research interest to date and, in general, the acronym MCCDMA is used to describe this scheme [8]. A more appropriate acronym often used for this scheme, which helps avoid the confusion with other MCCDMA schemes, is OFDM-CDMA.

In all previous MCCDMA schemes, the symbols of each user were first spread using a spreading sequence according to DSSS, and then different techniques were used to send the resulting chips over the channel using multicarrier transmission. There is one MCCDMA scheme, however, proposed by Vandendorpe [12] and termed *multitone CDMA* (MTCDMA), which reverses this order. According to this scheme, first a MCM block symbol is formed using N symbols by each user, and then this signal is spread in the time domain by multiplying it with the spreading sequence. The idea behind this scheme is that for a given data rate and chip rate, the duration of each symbol can be much larger than in a corresponding DSCDMA system because of the effect of MCM. This longer duration, in turn, is translated in longer DSSS spreading sequences and higher processing gain. This approach has a potential drawback compared to other MCCDMA systems since it needs a RAKE receiver or some form of equalization at the receiver. The concept of the transmitter structure for this scheme is depicted in Fig. 7.

3. OFDM-CDMA

OFDM-CDMA has attracted the largest research interest to date among all MCCDMA systems. It is a scheme that offers a large degree of flexibility in system design, high bandwidth efficiency, and good performance with acceptable complexity. OFDM-CDMA is more appropriate for the forward link (base station to mobile users) of wireless systems, since its performance is best when all

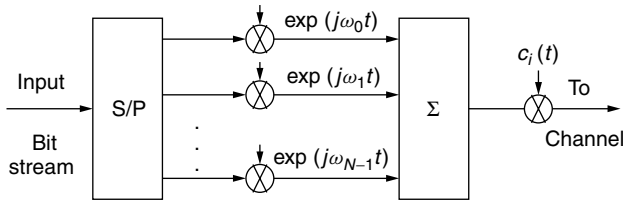


Figure 7. Vandendorpe MCCDMA scheme: transmitter structure.

users transmit in a synchronous manner. We will describe in more detail an OFDM-CDMA multiple-access system where N_u users are transmitting simultaneously in a synchronous manner using Walsh–Hadamard orthogonal codes of length N_s. Therefore up to N_s users can transmit at the same time. The nth MC block symbol (of duration T_b) for user i is formed by taking μ symbols $b_i^1(n), \ldots, b_i^\mu(n)$ in parallel, spreading them with the user's spreading sequence $\mathbf{c}_i = [c_{1,i} \ldots c_{N_s,i}]^T$, $c_{j,i} = \pm 1$, performing frequency interleaving, and placing the resulting $u_i^1(n), \ldots, u_i^{\mu N_s}(n)$ chips into the $N \geq \mu N_s$ available subchannels, each having width $\Delta f = 1/T_b$, by using an IFFT of size N. In practice N is larger than the number of subchannels μN_s required for the tranmission of the data in order to avoid frequency aliasing after sampling at the receiver. For that reason the data vector at the input of the IFFT is padded with zeros at its edges so that the $(N - \mu N_s)$ unmodulated carriers are split in both sides of the useful spectrum. The function of the identical frequency interleavers is to ensure that the N_s chips corresponding to each of the μ symbols are transmitted over approximately independently fading subchannels. As mentioned previously, this is possible only if μ/T_b is larger than the coherence bandwidth $(\Delta f)_c$ of the channel. After performing a parallel to serial conversion, a guard interval is added in the form of a cyclic prefix, and the signals of all the users are added and transmitted through the channel. The block diagram of this transmitter is depicted in Fig. 6. For simplicity we concentrate on only one of the μ symbols that each user transmits and we consider binary symbols $b_k(n) = \pm 1, k = 1, \ldots, N_u$ forming the data vector $\mathbf{b}(n) = [b_1(n), \ldots, b_{N_u}(n)]^T$, where n is the time index denoting the nth symbol interval. The transmitted signal during the nth MC block symbol period can be approximately written as follows:

$$s(t) = \sum_{k=1}^{N_u} \sum_{l=1}^{N_s} \sqrt{E_c} c_{l,k} b_k(n) e^{j[2\pi l(t-nT_G)/T_b]} \tag{4}$$

where $t \in [nT, (n+1)T]$, $T = T_b + T_G$, T_G is the guard interval chosen to be at least equal to the delay spread T_m of the channel, and E_c is the energy per chip.

Even in a frequency-selective, multipath fading channel, the fading of the narrow subchannels is approximately flat and is described by multiplicative complex channel coefficients $h_l(n), l = 1, \ldots, N_s$, which are samples of the channel frequency response at the center frequency f_l of the lth subchannel at $t = nT$. Because of the frequency interleaving function, the channel complex coefficient processes are approximately independent. Because of the existence of a guard interval with duration at least equal to the channel's delay spread, there is no intersymbol interference, and the signal received by user i can be approximately described by the following equation:

$$r(t) = ! \sum_{k=1}^{N_u} \sum_{l=1}^{N_s} \sqrt{E_c} h_l^{(i)}(n) c_{l,k} b_k(n) e^{j[2\pi l(t-nT_G)/T_b]} + \eta(t) \tag{5}$$

where $t \in [nT, (n+1)T]$, $h_l^{(i)}(n)$ are the complex channel coefficients that describe the channel between the transmitter and the user i, and $\eta(t)$ is the AWGN. At the receiver, the signal is sampled at a rate N/T_b, the samples that correspond to the cyclic prefix are discarded, an FFT of size N is performed, and frequency deinterleaving takes place. The vector $\mathbf{r}(n) = [r_1(n), \ldots, r_{N_s}(n)]^T$ at the output of the deinterleaver is given in matrix notation by the following equation:

$$\mathbf{r}(n) = \sqrt{E_c}\mathbf{H}(n)\mathbf{C}\mathbf{b}(n) + \boldsymbol{\eta}(n) \tag{6}$$

where $\mathbf{H}(n) = \text{diag}\{h_1(n), \ldots, h_{N_s}(n)\}$, matrix $\mathbf{C} = [\mathbf{c}_1 \mid \cdots \mid \mathbf{c}_{N_u}]$ is the $N_s \times N_u$ matrix whose columns are the spreading sequences of the users, $\mathbf{b}(n)$ is the data vector of the users, and $\boldsymbol{\eta}(n) = [\eta_1(n), \ldots, \eta_{N_s}(n)]^T$ is a vector containing zero mean, uncorrelated complex Gaussian noise samples, with variance $2\sigma^2$. The block diagram of an OFDM-CDMA receiver is depicted in Fig. 8.

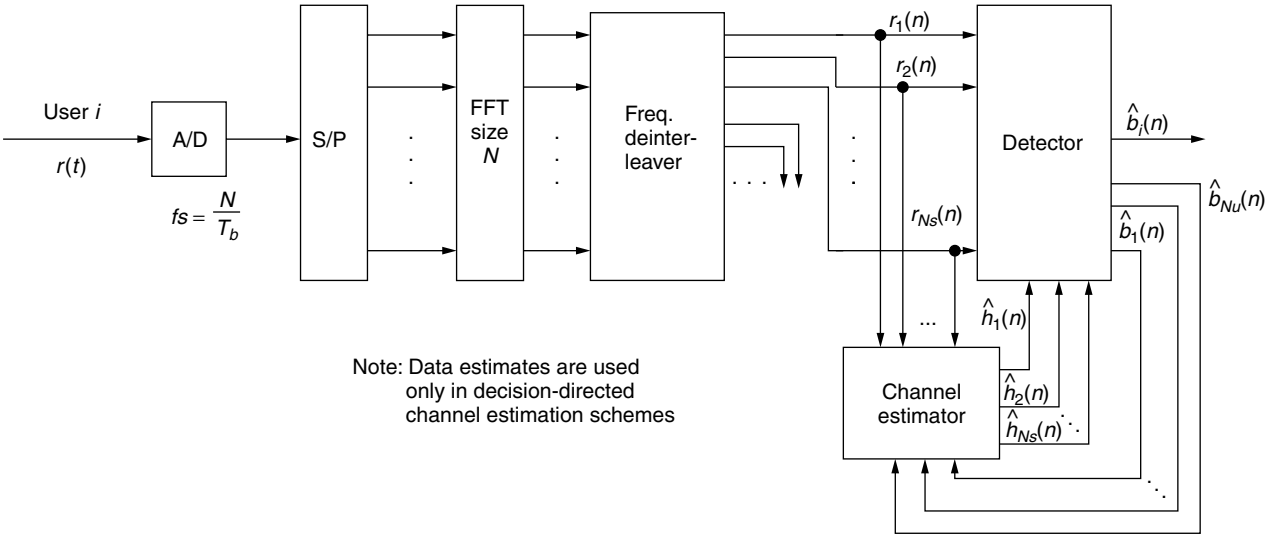


Figure 8. OFDM-CDMA receiver block diagram with channel estimation.

In the special case of an AWGN or flat-fading channel the channel coefficient matrix $\mathbf{H}(n)$ is within a constant the identity matrix. The optimal estimate of symbol $b_j(n)$ of user j is then obtained by correlating the received signal vector $\mathbf{r}(n)$ with the spreading sequence of user $\mathbf{c}_j$. This detector is also referred to as an *equal-gain combining* (EGC) detector. In this case, because of the orthogonality of the spreading codes used, there is no multiuser interference (MUI) and the performance is limited only by the AWGN. On the other hand, in the special case of a single-user system operating in a frequency-selective channel, the optimal symbol estimate is obtained by correlating the received signal vector $\mathbf{r}(n)$ with the spreading sequence of the user weighted by the complex conjugates of the channel coefficients. This detector is referred to as maximal ratio combining (MRC) detector. In the general case, however, of multiuser OFDM-CDMA systems operating over frequency-selective channels, the effect of the channel as demonstrated by matrix $\mathbf{H}(n)$ is to destroy the orthogonality among users and the performance is limited by the presence of both MUI and AWGN. The performance of both the EGC and MRC detectors in this case deteriorates significantly as the number of users increases and becomes unacceptable, even for relatively low system loads [13,14].

The optimal detector in the general case is the maximum-likelihood detector (MLD) [14,15]. This detector, however, may be impractical in many cases because its complexity grows exponentially with the number of users. For these reason, other low-complexity suboptimal detectors have been proposed, which attempt to combine good performance with reasonable implementation complexity. The simplest of such detectors, which corresponds to zero-forcing equalization, is termed *orthogonality restoring detector* (ORC). This detector inverts the effect of the channel by inverting matrix $\mathbf{H}(n)$, thus completely eliminating MUI. This causes, however, excessive noise enhancement due to the inversion of channel coefficients with very low magnitudes, which renders the performance of the ORC detector unacceptable [8,14]. More appropriate suboptimal detectors attempt to invert the effect of the channel without causing excessive noise enhancement. Examples of such detectors include the MMSE detector [10,14], the thresholded orthogonality restoring combining (TORC) detector [13,16] and the multistage or iterative OFDM-CDMA detector [15,16].

Most of the OFDM-CDMA detectors above require knowledge of the channel coefficient matrix $\mathbf{H}(n)$. For this reason a channel estimator is usually part of the receiver structure as shown in Fig. 8. The performance of practical OFDM-CDMA detectors operating in fast-fading channels deteriorates as a result of channel estimation errors. Examples of proposals for OFDM-CDMA detectors with channel estimation include pilot-symbol-based detectors [17] and decision-directed adaptive detectors [18].

4. COMPARISONS AND DISCUSSION

Few studies have been conducted to compare the performance of the different MCCDMA schemes. A comparison presented by Hara and Prasad [8] comparing OFDM-CDMA, MCDSCDMA, and MTCDMA shows that an OFDM-CDMA system with MMSE detection has the potential to outperform other schemes, with reasonable implementation complexity. The advantage inherent in OFDM-CDMA is higher frequency diversity. However, in designing an OFDM-CDMA system, calculation of the length of the spreading sequences, which determines the number of subchannels used to transmit each symbol, should be based on the maximum available frequency diversity as expressed by the ratio of total bandwidth over channel coherence bandwidth. If, for example, this ratio is in the order of 8, transmitting each symbol over $N_s \geq 64$ subchannels should not bring significant performance benefits. MCDSCDMA schemes using orthogonal carriers and introducing maximum frequency diversity based on the above ratio should demonstrate similar performance and capacity. On the other hand, MCDSCDMA schemes using nonoverlapping subchannels [6] have inherently lower bandwidth efficiency because of the necessary guardbands between adjacent subchannels. Other factors may favor the use of such schemes, however, such as the lack of intercarrier interference and ease of frequency synchronization. As a general comment, one has to be cautious before declaring one MCCDMA scheme as the "best" because implementation factors and the operational environment (e.g., the nature of interference sources and of the wireless channel) may impact the performance of each scheme differently and significantly.

Another topic that has attracted the interest of many researchers is the comparison between DSCDMA and MCCDMA. It has been reported [19] that there are cases where OFDM-CDMA systems with MMSE or MLD detectors significantly outperform DSCDMA systems, especially in the forward link of heavily loaded systems. Another argument in favor of OFDM-CDMA is that in practice RAKE receivers have a small number of fingers and thus may not be able to use all available signal energy, which may give some OFDM-CDMA systems with long symbol duration a certain advantage [8]. However, if all design aspects are taken into consideration (spreading with long random sequences, implementation issues such as channel estimation and carrier synchronization, channel coding), properly designed orthogonal MCCDMA and DSCDMA systems might plausibly demonstrate similar performance and system capacity [20]. Again, implementation issues and the operational environment can significantly affect the performance of each system and must be considered carefully before a system design is selected.

Finally, there has been research on the issue of comparing OFDM-CDMA with multiuser OFDM (MOFDM) systems [15,21,22]. OFDM can support multiuser systems without the use of spreading codes, either by assigning each user a subset of the available subchannels or by allowing only one user to transmit an MCM block at a given time as in TDMA systems. In general, OFDM-CDMA systems have higher complexity than do corresponding MOFDM systems; therefore their use is justified only when superior performance can be achieved. It was shown that in uncoded systems OFDM-CDMA significantly outperforms MOFDM because of the inherent

higher frequency diversity [21,22]. When channel coding is used, however, the performance gain due to increased frequency diversity introduced by frequency-domain coding is so much higher in MOFDM systems that their performance becomes similar to the more complex coded OFDM-CDMA [21,22]. Coded OFDM-CDMA systems can still outperform MOFDM at the expense of using more complex detectors (e.g., the MLD), especially for higher coding rates $>1/2$ [15]; however, coded MOFDM systems with lower coding rates ($<1/2$) have similar performance [15], and use of the more complex OFDM-CDMA is not justified in this case.

BIOGRAPHY

Dimitris N. Kalofonos received the Dipl.Ing. degree from the National Technical University of Athens (NTUA), Athens, Greece, in 1994, and the M.Sc. and Ph.D. degrees in electrical engineering from Northeastern University, Boston, Massachusetts in 1996 and 2001, respectively. From 1993 to 1994 he was with the Microwave Systems department in Intracom S.A., Athens, Greece, where he worked on DSP design for wireless systems. From 1996 to 2000 he was with the Wireless Systems department of GTE/Verizon Laboratories, Waltham, Massachusetts, where he conducted research on performance modeling of 2G and 3G CDMA cellular networks. From 2000 to 2001 he was with the Mobile Networking Systems Department of BBN Technologies, Cambridge, Massachusetts, working on adaptive waveform design for mobile ad hoc networks. He is currently a Senior Research Engineer at the Communication Systems Laboratory of Nokia Research Center in Boston, where he is conducting research on pervasive networking and mobile Internet technologies. His interests include wireless personal-area and local-area networks (PAN, LAN), ad hoc networks, and wireless integrated services networks. Dr. Kalofonos is a member of the Technical Chamber of Greece and a registered engineer in Greece.

BIBLIOGRAPHY

1. J. Bringham, Multicarrier modulation for data transmission: An idea whose time has come, *IEEE Commun. Mag.* 5–14 (May 1990).
2. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, 1995.
3. V. DaSilva and E. Sousa, Performance of Orthogonal CDMA codes for quasi-synchronous communication systems, *Proc. IEEE Int. Conf. Universal Personal Communications* (*ICUPC'93*), 1993, Volume 2, pp. 995–999.
4. Q. Chen, E. Sousa, and S. Pasupathy, Performance of a coded multi-carrier DS-CDMA system in multi-path fading channels, *Wireless Pers. Commun.* **2**: 167–183 (1995).
5. S. Kondo and L. Milstein, On the use of multicarrier direct sequence spread spectrum systems, *Proc. IEEE MILCOM*, 1993, Vol. 1, pp. 52–56.
6. S. Kondo and L. Milstein, Performance of multicarrier DS-CDMA systems, *IEEE Trans. Commun.* **44**(2): 238–246 (Feb. 1996).
7. E. Sourour and M. Nakagawa, Performance of orthogonal multicarrier CDMA in a multipath fading channel, *IEEE Trans. Commun.* **44**(3): 356–367 (March 1996).
8. S. Hara and R. Prasad, Overview of multicarrier CDMA, *IEEE Commun. Mag.* 126–133 (Dec. 1997).
9. K. Fazel, Performance of CDMA/OFDM for mobile communication systems, *Proc. 2nd IEEE Int. Conf. Universal Personal Communications* (*ICUPC*), 1993, pp. 975–979.
10. A. Chouly, A. Brajal, and S. Jourdan, Orthogonal multicarrier techniques applied to direct sequence spread spectrum CDMA systems, *Proc. IEEE Global Communications Conf.* (*GLOBECOM'93*), 1993, pp. 1723–1728.
11. N. Yee J. Linnartz, and G. Fettweis, Multi-carrier CDMA in indoor wireless radio networks, *Proc. PIMRC*, Yokohama, Japan, 1993, pp. 109–113.
12. L. Vandendorpe, Multitone spread spectrum multiple access communications system in a multipath Rician fading channel, *IEEE Trans. Vehic. Technol.* **44**(2): 327–337 (May 1995).
13. T. Muller, H. Rohling, and R. Grunheid, Comparison of different detection algorithms for OFDM-CDMA in broadband Rayleigh fading, *Proc. IEEE Vehicular Technology Conf.* 1995, 835–838.
14. S. Kaiser, Analytical performance evaluation of OFDM-CDMA mobile radio systems, *Proc. 1 European Personal and Mobile Communications Conf.,* (*EPMCC'95*), Bologna, Italy, Nov. 1995, pp. 215–220.
15. S. Kaiser, *Multi-Carrier CDMA Mobile Radio Systems—Analysis and Optimization of Detection, Decoding, and Channel Estimation*, Ph.D. thesis, VDI-Verlag, Fortschittberichte VDI, Series 10, No. 531, 1998.
16. D. N. Kalofonos and J. G. Proakis, Performance of the multistage detector for a MC-CDMA system in a Rayleigh fading channel, *Proc. IEEE Global Communications Conf.* (*GLOBECOM'96*), Nov. 1996, Volume 3, pp. 1784–1788.
17. S. Kaiser and P. Hoeher, Performance of multi-carrier CDMA with channel estimation in two dimensions, *Proc. IEEE Symp. Personal Indoor and Mobile Radio Communications* (*PIMRC'97*), 1997.
18. D. N., Kalofonos, M. Stojanovic, and J. G. Proakis, Analysis of the impact of channel estimation errors on the performance of a MC-CDMA system in a Rayleigh fading channel, *Proc. IEEE Communications Theory Mini Conf.* (*CTMC*) *in Conjunction with GLOBECOM'97*, Nov. 1997, 213–217.
19. S. Kaiser, OFDM-CDMA vs DS-CDMA: Performance evaluation for fading channels, *Proc. IEEE Int. Conf. Communications*, 1995, pp. 1722–1726.
20. S. Hara and R. Prasad, Design and performance of multicarrier CDMA systems in frequency-selective Rayleigh fading channels, *IEEE Trans. Vehic. Technol.* **48**(5): 1584–1595 (Sept. 1999).
21. J.-P. Linnartz, Performance analysis of synchronous MC-CDMA in mobile Rayleigh channel with both delay and doppler spreads, *IEEE Trans. Vehic. Technol.* **50**(6): 1375–1387 (Nov. 2001).
22. C. Ibars and Y. Bar-Ness, Comparing the performance of coded multiuser OFDM and coded MC-CDMA over fading channels, *Proc. IEEE Global Communications Conf.* (*GLOBECOM'01*), 2001, 881–885.

MULTICAST ALGORITHMS

AIGUO FEI
MARIO GERLA
University of California at Los Angeles
Los Angeles, California

1. GROUP COMMUNICATION AND MULTICAST

Group or *multipoint communication* [12] refers to the type of communication in which information is exchanged among multiple (more than two) communication entities simultaneously. Many applications involve multipoint communication in nature, including videoconferencing, distance learning, distributed database synchronization, real-time distribution of news or stock quote, and multiplayer Internet gaming. On the other hand, as the fundamental method of telecommunication or computer communication, point-to-point communication is information exchange between two entities (although there may be many entities in between to help transport information). Modern communication networks have been very successful and efficient in supporting this type of communication service, while support for group communications is more a recent development and is likely to take many more years to mature.

Providing multipoint communication services is a multifacet problem: (1) addressing, group management, and membership management — how to identify and manage different communication groups and how to identify and manage members of a group; (2) session management — how to initiate a group communication session and how to control information transmission from one member to the others; (3) traffic control (e.g., not to let a sender overflow the network or a receiver); (4) reliability or data integrity — some applications may require any data sent from a source to be delivered to all other participants reliably, while some other applications may not have such requirement; and (5) data or information distribution — how data are distributed from their source to other participants.

The focus of this article is the last aspect of the problem: how to build a delivery structure to disseminate data for a communication group.

1.1. Group Communication in Shared-Medium Networks

There two fundamentally different types of networks: point-to-point networks and shared-medium networks. In a point-to-point network, every "link" in the network connects two stations and any data transmitted over that link by one station are received by and only by the station at the other end. In a shared-medium network (Fig. 1), any data transmitted by one station will be received by all others in the same network (or by those within a certain range as in wireless networks), although some stations may discard the data if they are not interested in them. Ethernet- or token-ring-based local-area networks (LANs) are examples of the shared-medium networks. Wireless LAN is another example. The nature of "shared medium"often limits this type of network to within a small range [i.e., LANs or metropolitan-area networks (MANs) at most].

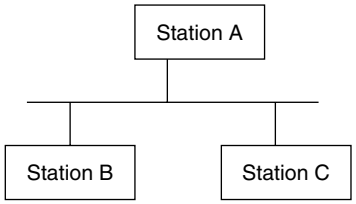


Figure 1. A shared-medium local-area network. Stations A, B, and C share a "same" wire; any data transmitted to the wire by one station will be seen by all others.

Here we discuss group communication in a shared-medium network using IEEE 802 LAN [31] as an example. In an IEEE 802 LAN network, each station can be identified by a unique 48-bit "individual" MAC (medium access control) address, as illustrated in Fig. 2. A data packet (i.e., usually called an "Ethernet frame" in an Ethernet network) targeted to a specific destination carries the address of the destination station. Some MAC addresses are "group" addresses that have the "group/individual" bit set to 1 (while the bit in an individual address is 0). One particular group address is called "broadcast" address, which is all 1s. Packets targeted for a specific group carry the corresponding group address (the broadcast address if targeted for all stations in the network). A station receives all packets transmitted in the network; however, except in some cases, it doesn't need to process every packet. The network interface device of a station can filter out packets except those that carry the station's unique address or addresses of groups that it is interested in. In some sense, group communication in a LAN comes (almost) "free"; a station can receive every packet transmitted — it needs to select only those in which it is interested. The set of groups a station is interested in is determined by the protocol layer above the MAC layer — we will discuss the case of IP.

In a more advanced switched Ethernet network [31] (or other type of switched LAN), the assumption that any packet transmitted by one station is received by all others is no longer true. However, the same address scheme is still used, and a switch delivers all packets with group

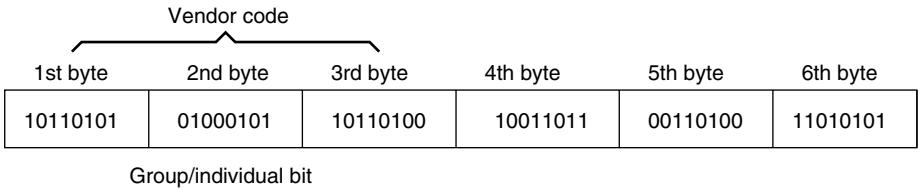


Figure 2. IEEE 802.11 MAC address. The higher 3 bytes are assigned to vendors to identify different vendors; a vendor then assigns a unique value of the lower 3 bytes to each network interface card it manufactures.

addresses to every other station in the network. Thus it makes no difference to a higher-layer protocol that utilizes the group communication capability of the IEEE 802 LAN.

1.2. Group Communication in Point-to-Point Networks

A point-to-point network consists of network nodes and point-to-point links connecting the nodes. Data transmitted by one node over a link are received only by the node at the other end of the link (i.e., a direct neighbor). Data destined to a single arbitrary node may travel a multihop path found by a routing protocol [19] to reach its destination. This type of network service is called *unicast*, and a point-to-point network can support it very efficiently. However, it takes more elaborate control to deliver data from one station to multiple destinations as required by group communications.

The simplest method to support multipoint communication is broadcasting. In this method, when a node receives a data packet, it sends the packet to every other neighbor except the one from which it receives the packet. This way, every packet is "broadcasted" from its source to all other nodes in the network. Conceptually broadcasting is very easy to implement; however, care must be taken to ensure that packets are "killed" somewhere so that they will not "loop" forever. This approach is extremely inefficient in network resource usage since there will be a lot of unwanted packets floating around in the network and bandwidth is wasted in transmitting them.

The second approach can be called a "naive unicast" approach, in which a source node sends a copy of the data to every other group member that is interested in receiving data from it, through unicast. This is illustrated in Fig. 3a; assuming that a group consists of nodes S, D, E, and F, among which S is the source node, when S wants to send data to the group, it sends a copy to each of them (D, E, and F). The third approach can be called a "server-based" unicast approach as illustrated in Fig. 3b; assume that node B is the server, S sends packets to B, and B forwards packets received to all other members (D, E, and F) through unicast. If another member node wants to send data to the group, it also sends the data to B, and B forwards it to everyone else. If there is only one single source, these two approaches would be the same if the server is placed at the source node. However, if any group member can be a traffic source, then the server-based approach has its advantages; every member node only needs to know what node is the server and doesn't need to know each other (which greatly simplifies group management). Clearly these two approaches are more efficient in terms of resource usage than broadcast.

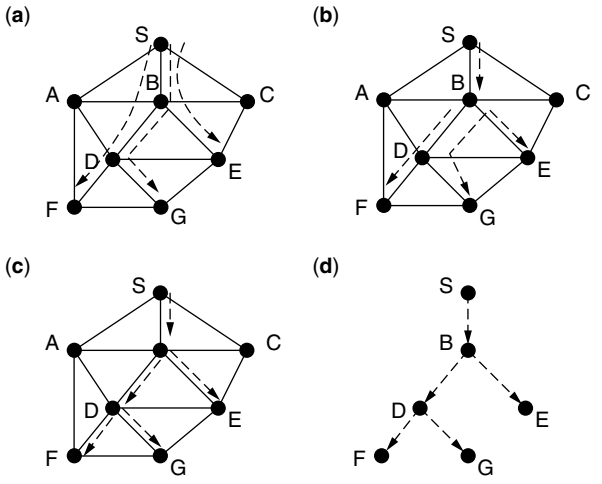


Figure 3. Different solutions for group communication: (**a**) naive unicast, (**b**) server-based unicast, (**c**) multicast, (**d**) a multicast tree.

Unicast-based approaches to group communication have some nice features: (1) they rely only on the unicast capability already provided by the network; no additional support is required for data delivery over the network (although other helper entities such as a server need to be introduced to help forward data among group members) and (2) they are conceptually easy to implement. Indeed, server-based solution is widely used to support group communication in the Internet today (Web-based chatroom, multiperson Internet gaming, etc.). In the public telephone network (including ISDN), services involving group communication [three-way conference call, multipoint videoconferencing using a multipoint videoconferencing control unit (MCU) [7]] are also implemented using unicast. However, they also suffer some clear drawbacks: scalability problems and resource efficiency. In both the naive unicast and server-based approaches, the scalability problem is a twofold problem: (1) a single node or a server has limited bandwidth to connect to the network, which limits the number of group participants that it can support; and (2) a single node or a server itself has limited local resource, which also limits the number of members it can support. The second disadvantages would only be seen in contrast of the multicast approach to group communication, which discussed next.

The fourth approach to group communication would be "multicast." This is illustrated in Fig. 3c; when node S wants to send data to the group, it sends a copy to node A, which forwards a copy to E and D, which then sends a copy to F and G. The paths taken by the data packets collectively form a tree delivery structure which is called a "multicast tree" as illustrated in Fig. 3d.

One advantage of the multicast approach is resource efficiency; for example, a data packet is forwarded only once over link ($B \rightarrow D$) to reach nodes F and G, but it is forwarded twice over that link (once by each unicast connection to F and G) in unicast-based approaches. Another advantage is better scalability with group size; if the network is large and a group has many members, any node in the multicast tree needs to forward a packet only to its neighbors in the tree (which are normally of a small number), while in unicast-based approaches, the source or center has to send many copies of a data packet to reach all other members.

Of course, the advantages of multicast are not achieved without a price — explicit extra support (besides unicast) from the network is required to do multicast forwarding. For example, when node B receives a packet from S, it must know that the packet is a multicast packet and that the packet should be forwarded to neighbors D and E (but

not *A* and *C*). This means that extra packet processing has to be done, and some extra state information regarding a multicast group has to be maintained at network nodes. The problem of constructing a multicast tree in a network and installing necessary state information is the multicast routing problem, which is accomplished by a multicast routing protocol, and at the center of that there is a multicast routing algorithm.

2. MULTICAST ROUTING IN THE INTERNET

A routing algorithm is part of a routing protocol. Before we go into any algorithm details, we give an overview of the multicast routing architecture in the Internet.

2.1. Overview

Conceptually the Internet can be modeled as a three-level hierarchy (see Fig. 4). At the lowest level, the computers (hosts) of end users are connected together to form a local-area network (LAN) such as an Ethernet network — this could be a single computer for a home user. Each LAN has a *designated router* that connects all the computers of that LAN to the Internet — the vast network consisting of all the computers interconnected together. For a home user, the designated router is normally the access router at the Internet service provider (ISP) side, which a home computer connects to through a dialup line or a cable/DSL modem. For business users, their LAN routers are also connected (e.g., through other routers within their own networks) to access routers at service providers.

Access routers of an ISP network are connected together through other routers to form an intradomain network, the second level. Some large corporations may also have multiple LANs connected together to form an intradomain network as well. At the highest level, different domains are interconnected together through border routers to form an interdomain network (i.e., a network of domains). Very often intradomain and interdomain networks are point-to-point networks. Although routers within a domain or border routers between domains are sometimes connected through a shared-medium network, they are often treated as a point-to-point network logically for routing purposes.

In the current IP multicast architecture, a multicast group is identified by an IP address of a special class — the class D IP addresses that have the first 4 bits as 1110 (thus a multicast address is within the range of 224.0.0.0–239.255.255.255). IP multicast follows a very simple model:[1] (1) a host that joins a group with a specific group address shall receive any packet sent to that group and (2) a packet sent to a group address from any host shall be received by all members of that group. Next we discuss how (1) and (2) would be implemented in IP networks.

2.2. Multicast at LAN Level

At LAN level, a host joins a multicast group by communicating with the designated router (DR) via the Internet Group Membership Protocol (IGMP [6,15]): (1) a host can send a message (membership report) to the DR to join a specific group; (2) the DR may periodically send query messages for a group, and a host can answer this query to express continuous interest in that group or silently drop the message to indicate that it is no longer interested in that group; or (3) a host may send "leave group" message to the DR to explicitly leave a group. A host operating system supporting multicast provides API (application programming interface) functions for applications to join or leave multicast groups.

The question now is how a host sends or receives multicast traffic. We use an Ethernet network as an example. To send IP packets for a group, a host just puts the group IP address as the destination address and sends it to the DR encapsulated in an Ethernet frame with DR's MAC address, which is the same as sending a unicast packet. A DR is responsible for sending it out to reach other group members (which we discuss next). A host doesn't need to join a group to send data to that group.

Receiving multicast packets is done through mapping of an IP multicast address to a MAC address. The IETF (Internet Engineering Task Force [1]) has a single 802 MAC address block (01-00-5E-00-00-00) with the lowest 24 bits assignable. A multicast IP address has 28 unique bits (the highest 4 bits are 1110), the lowest 23 bits of

[1] Some applications may enforce some kind of access control; thus these two rules may not always apply. Initially IP multicast was designed to support only the model at the network layer; thus such access control has to be implemented elsewhere. More recently, source-specific control was introduced into IP multicast [6]. Now, say, a host can join a group while specifying that it is interested only in receiving packets from a list of sources.

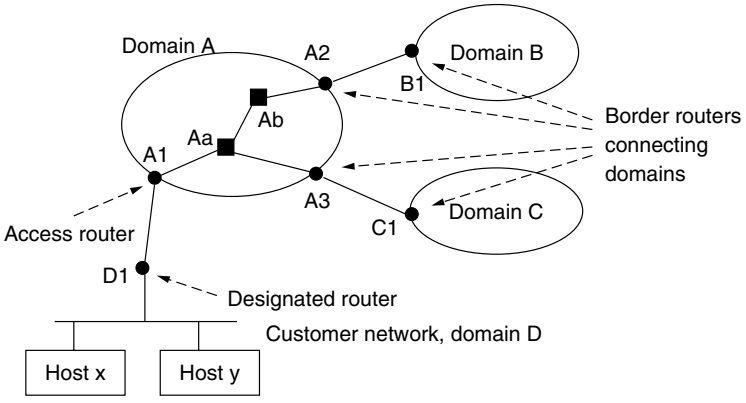


Figure 4. Internet hierarchy. The routers shown here together form a multicast tree.

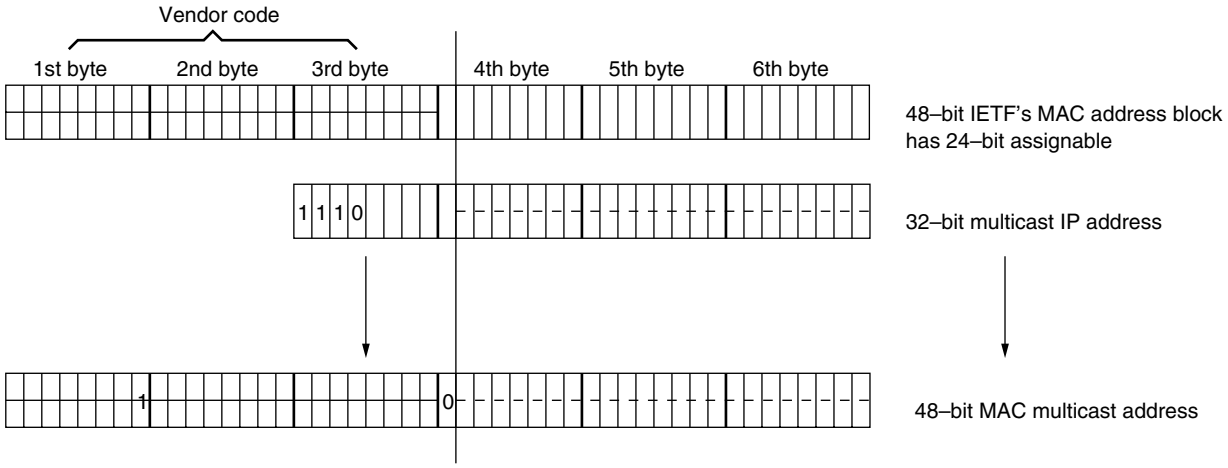


Figure 5. Map of a multicast IP address to a multicast MAC address. Half of IETF's address block (the 24th bit is 0) is used for multicast; the other half (the 24th bit is 1) is reserved.

that 28 bits are mapped into the lowest 23 bits of IETF's address space to form a MAC multicast address for that group, as illustrated in Fig. 5. When a host joins a group, it informs the network interface card to listen to the mapped MAC address. A DR sends multicast packets received from external routers for that group to the LAN network by encapsulating them in an Ethernet frame with the mapped address. One may note that, since an IP multicast address has 28 unique bits, there may be two or more different groups mapped into a single MAC address at the same time within a LAN. This won't be a problem since the original multicast address is carried in the IP header and the receiving host can just discard packets for groups from which it does not intend to receive data.

2.3. Multicast over Wide-Area Networks

Wide-area networks (WANs) are often point-to-point networks. IP multicast utilizes a tree structure to deliver multicast packets across WANs as described earlier. A tree consists of designated routers that have group members in their subnets and other intermediate routers that help transport multicast traffic in between. For example, as shown in Fig. 4, a multicast tree may consist of router $D_1, A_1, A_2, A_3, A_a, A_b, B_1, B_2$, and some other DR routers in domains B and C.

Once a router is in the tree for a specific group, it uses forwarding-state information to determine how to forward multicast packets received. In source-tree based multicast routing protocols, forwarding-state entries are intended for per group/source. For example, a forwarding entry at a router is like *group:g / source:s, expected in — interface:I, out — interface*(s):$O_1 \cdots O_n$. When this router receives a packet with destination address g and source address s from interface I, it will send the packet out to interface $O_1, \ldots, O_n$. In group shared-tree protocols, there is one entry for one group. For example, a forwarding entry will be *group:g, list of interfaces*: $I_1, I_2, \ldots, I_n$. A packet with destination address g received from interface I_{in} will be sent out to all other interfaces in the list except I_{in}.

Multicast routing protocols determine how a multicast tree is constructed. Matching the Internet hierarchy, routing protocols are divided into intradomain protocols and interdomain protocols. Normally a domain [often called an *autonomous system* (AS)] is controlled by a single administrative entity and can run an intradomain multicast routing protocol of its choice. An interdomain multicast routing protocol is deployed at border routers of a domain to construct multicast trees connecting to other domains. A border router capable of multicast communicates with its peer(s) in other domain(s) via interdomain multicast protocols and routers in its own via intradomain protocols, and forwards multicast packets across the domain boundary.

3. STEINER TREE PROBLEM

Multicast routing algorithms are closely related to the *Steiner tree problem* [20] in graph theory [21]. The Steiner tree problem is related to the *minimum spanning-tree* problem, which can be stated as follows. Given a connected graph $G(V, E)$, where V is a set of vertices(nodes) and E is a set of edges, and for each edge $e \in E$ it has a weight $w(e)$, a minimum spanning tree is a tree $T(V, E')$ $(E' \subset E)$, which contains all the nodes in G and its weight $W(T) = \sum_{e \in T} w(e)$ is minimal of all possible spanning trees. The minimum-weight spanning tree that covers a subset of V is a Steiner tree: tree $T(V', E')$ with minimal weight $W(T)$ for a given $V' \subset V$.

There are two classical algorithms for the minimum spanning-tree problem [33]: (1) *Prim's algorithm*, which starts with an arbitrary node and grows the tree by repeatedly adding the minimum-weight edge that connects an in-tree node to a node that is not yet in the tree until all nodes are connected; and (2) *Kruskal's algorithm*, which initially has each node as a separate tree and then constructs the Steiner tree through merging them into one by repeatedly adding the minimum-weight edge that connects two trees (into one) without creating a cycle.

The Steiner tree problem is NP-hard in general, which means that the time it takes to find exact solutions is exponential regarding graph size. A number of heuristic algorithms have been proposed [16,20] to find approximate

solutions in P time. A shortest-paths heuristic proposed by Takahashi and Matsuyama (TM algorithm) [20,34] is a simple heuristic that has a proven good performance bound. Based on a greedy strategy, it starts with a subtree T consisting of a single node arbitrarily chosen from V'. The tree T grows by adding the node from V' that has the shortest distance to nodes in T and is not covered by T, one by one until all nodes of V' are present in T.

In another heuristic proposed by Kou, Markowsky, and Berman (KMB algorithm) [20], first a complete graph $G'(D, F)$ is constructed with nodes $D = V'$, using the cost of the shortest path from i to j in G as the cost for edge $e_{ij} \in F$. Then the minimum spanning tree T' of G' is built. The spanning tree in G covering V' is obtained by replacing any edge (i, j) in G' with the shortest path $p(i \to j)$ in G.

In multicast routing, the goal is to construct a multicast tree that covers all the designated routers ("terminal nodes") for a multicast group so that all members can receive data sent to that group. On the other hand, one motivation for multicast is network resource (i.e., bandwidth) efficiency. Therefore we want to build a multicast tree that minimizes resource usage. For that purpose, we can assign to each link a weight that represents the cost to transport a unit of data over that link (i.e., the cost to use the bandwidth resource). Thus the problem of constructing a multicast tree that minimizes resource usage is to find a Steiner tree that covers all the terminal nodes of a group.

4. INTRADOMAIN ROUTING PROTOCOLS AND ALGORITHMS

From our presentation of the two Steiner tree algorithms, we can see that there are three elements in multicast routing: (1) network information [network nodes and how they are connected, i.e., $G(V, E)$], (2) group membership information (the set of nodes that we need to cover, V'), and (3) an algorithm to construct the tree.

None of the existing multicast protocols actually uses the TM algorithm or KMB algorithm in Section 3 because of their two requirements, as follows: (1) complete network topology information is needed [i.e., $G(V, E)$] and (2) all terminal nodes must be known in advance (i.e., V'). These two requirements make it difficult or impossible to implement either algorithm in a distributed routing environment like the Internet. At the same time, the IP multicast model assumes dynamic membership (i.e., members can join or leave at any time) and supports nonmember sending (i.e., a host can send data to a group without joining it). The implication of requirement 2 is that whenever there is a membership change, a multicast tree must be recomputed. To support this, it is more desirable to have a routing algorithm that would construct a multicast incrementally as members join a group and introduce minimal modification to the existing tree when a member leaves the group. At the same time, nonmember sending has to be supported. Nevertheless, optimal Steiner trees produced by good approximate algorithms are often used to compare with those constructed by IP multicast protocols in performance studies.

4.1. MOSPF: Shortest-Path Tree

MOSPF [27,28] is an extension of the Open Shortest Path First (OSPF) protocol to support multicast routing in an intradomain environment. OSPF is a link-state routing protocol [19,29], at the core of which there is a distributed and replicated link-state database at each router in an AS. This database is like a map of the network describing all routers within and their interconnections. It is constructed and updated through flooding of link-state advertisements (LSAs). Each LSA describes the links of a node to its neighbors (containing additional information such as *cost* of a link). For example, an LSA from router S may look like $\{S \to A: 2, S \to B: 1, S \to C: 3\}$, where the numbers are cost of the links, respectively. We also assume symmetric links here — for instance, there is also a link $(A \to S)$ with cost of $(A \to S)$ = cost of$(S \to A)$. Each router gathers LSAs from all other nodes in the network and constructs a complete topology of the network. On the basis of that topology, a router computes a routing table based on which to forward data packets. For example, an entry in the routing table could be like {131.169.96.0/24,[2] interface I}, which tells the router to forward a data packet received with destination address in the range from 131.169.96.0 to 131.169.96.255 to the interface I (which connects to a particular neighbor, the next hop for those packets). The algorithm used in OSPF is Dijkstra's algorithm [25], which computes the shortest paths from one node to all others. For example, assume that router F is the designated router for a subnetwork 131.169.96.0/24; node S computes the shortest path to F: $(S \to B \to D \to F)$; this tells S to forward packets destined to network 131.169.96.0/24 to neighbor B.

MOSPF extends OSPF to support multicast routing: (1) a new LSA is introduced to propagate group membership information, then (2) a router computes a multicast tree and determines the forwarding entry when it receives multicast packets for a group from a specific source. For example, when a host in subnet 131.169.96.0/24 sends an IGMP message to F to join a group g_x, F will flood the network with an LSA saying that subnetwork 131.169.96.0/24 is now a member of group g_x. We also assume that there are group members in router E's and G's subnets. Now a host in router S's subnet is a source sending traffic to group g_x. When S receives the first packet from that host to g_x, it computes all the shortest paths to all other nodes; then it recursively prunes routers that don't have any group members to get a multicast tree. This is illustrated in Fig. 6, where all nodes shown are routers; hosts are not shown since they are not involved in the routing process. From that tree, S knows that it should forward the packet to B. When B receives the packet, it computes a tree using its database of the network topology — the database is synchronized and the tree computed will be the same as S's, and B determines that it should forward the packet to D and C. Forwarding entries are cached. So when S receives the next packet from the same source to g_x, it knows how

[2] The number 24 means that the highest 24 bits are significant bits fixed as 131.169.96, while the lowest 8 bits can vary from 0 to 255 in value.

to forward it without computing the tree again. A cached entry also has a lifetime — it expires after a while and a router will compute it again; this way, a router doesn't waste bandwidth to forward multicast packets of group g_x to neighbors that no longer lead to group members. For example, after a while E is no longer a member of group g_x and wouldn't send out an LSA saying that it is; when the forwarding entry expires, B recomputes the tree and knows that it no longer needs to forward packets destined for g_x to E.

4.2. DVMRP: Broadcast and Prune

The Distance Vector Multicast Routing Protocol (DVMRP) [30] was developed to support multicast routing in an intradomain environment where a routing protocol of the *distance vector* protocol family is deployed. In the Internet, the Routing Information Protocol (RIP) [18,19] is a distance vector protocol that was widely deployed and used for intradomain routing before OSPF was developed. Instead of flooding link-state information, nodes exchange "distance" (to all other nodes) information with neighbors in RIP, and a distributed version of the Bellman–Ford shortest-path algorithm [9] is implemented for a node to figure out how to reach any other node. Unlike OSPF, in which each node maintains a complete topology of the network, in RIP a node only knows what next hop to go to reach a destination. For example, in the network shown in Fig. 6a, S knows that it should forward packets destined for node F's subnet to its neighbor B, while F knows that it should forward packets destined for node S's subnet to D. Similarly, B knows that it should forward packets destined for node S's subnet to S.

DVMRP employs a "broadcast and prune" approach to construct multicast trees. In MOSPF, when a node receives a multicast packet and doesn't know how to forward it [i.e., there is no forwarding entry for the (source, group) pair], it computes a shortest-path tree (rooted at the source) using the topology database and determines the forwarding entry. In DVMRP, when a node receives a multicast packet with source address s and destination (group) address g_x and it knows nothing about it, it does the following: (1) it checks whether the packet is received from the interface to which it normally forwards packets destined for s and (2) if not, the packet is dropped; otherwise the packet is forwarded to all other interfaces except the one from which the packet is received. For example, in Fig. 6, let's assume that the source host with IP s is in router S's subnet; when D receives a packet of (s, g_x) from node A, it will drop it because D always forwards packets destined for s to B instead of A; however, if that packet is received from B, D will forward it to nodes A, F, G, and E. For F and G, this is great, because both F and G have group members in their subnets. However, A and E don't have member in their subnets. What A or E does is to send a "prune" message to D to tell D not to forward multicast packets (s, g_x) to them anymore. D will remember that by adding a cache entry and will forward packets (s, g_x) to F and G only in the future. Similarly, when B receives a multicast packet with (s, g_x) from S, it forwards it to A, D, E, and C. After it receives prune messages from A and C, it remembers it and no longer forwards packets to them. But D will not send a prune message to B because D has nodes F and G at downstream and these nodes don't tell D that they don't want the packets (which translates into "they want those packets"). This way, eventually a multicast tree rooted at S and reaching F, G, and E is built. The cached prune information is periodically cleared. For example, after a while B will "forget" that it doesn't need to forward multicast packets (s, g_x) to A and C and again forwards them to all neighbors except S; then A and C will send prune messages again to "opt out" of the tree.

In DVMRP, a multicast tree is a reverse shortest-path tree rooted at the source compared with the shortest-path tree in MOSPF. In MOSPF, the path from source S to a destination F on the tree is the shortest path from S to F; in DVMRP, however, the path from F to S on the tree is the shortest path but the path from S to F is not necessarily the shortest (thus the term *reverse shortest-path tree*). In our example, they happen to be the same because we assume every link to be symmetric. In the modern Internet, "asymmetric" routing may happen (i.e., packets from F to S may travel a different path than packets from S to F) and the reverse shortest-path tree may not always be the same as the shortest-path tree.

4.3. PIM-SM: Reverse Shortest-Path Tree

There are some limitations with DVMRP and MOSPF. DVMRP is not very efficient because of the periodical flooding of multicast packets. This also leads to a scalability problem — when the network grows larger or when the number multicast groups grows larger, the efficiency problem becomes more severe. MOSPF also has a scalability problem; when the network size and/or the number of groups grows large, the processing requirement to compute trees may become excessive for a router. Another limitation is they are only

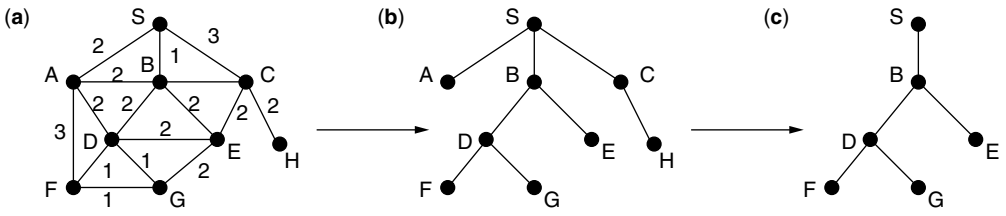


Figure 6. Construction of a shortest-path multicast tree: (**a**) network with link cost; (**b**) shortest-path tree from node S to all others; (**c**) get the multicast tree by recursively pruning nodes not interested in the group; remove node A and link (S, A), remove node H and link (C, H), remove node C and link (S, C).

applicable to intradomain routing. For these reasons, the Protocol Independent Multicast–Sparse Mode (PIM-SM) protocol [10,11,13] was developed.

PIM-SM also builds a reverse shortest-path tree. However, it is significantly different from DVMRP: (1) the tree is rooted at a central router called the *rendezvous point* (RP); (2) instead of broadcast-and-prune, it employs an explicit join mechanism to construct the tree. In PIM-SM, every multicast group has a central router (RP) responsible for that group. The RP is identified by its own unicast IP address and the RP for a group can be obtained or made known through a query or advertisement mechanism. Beyond that, similar to DVMRP, PIM-SM assumes that a router can provide the next hop information for any destination — this is easily and readily supported by any unicast routing protocol (this is why it is called protocol-independent; it doesn't rely on any specific unicast routing protocol). When a designated router (e.g., router S in Fig. 6) receives a membership report for a group from a host in its subnet, it sends a PIM-SM join request toward the RP router of the group. A PIM-SM capable (i.e., understands and supports PIM-SM) router will look up the next hop for the RP's address (a unicast address) and forward it to the next PIM-SM-capable router. That join request will stop when it reaches either (1) a router that already has multicast state (i.e., forwarding entry) for that group or (2) the RP router of the group. In either case, intermediate routers will install forwarding entries for the group and a new branch connecting to the new member is established. As an example, assume that router B is the RP for group g_x. When F first joins the group, it sends a join message destined for B to D (D is the next-hop router to which F should forward packets to reach B); D doesn't have a forwarding entry for g_x yet, it will create one and forward the request to B, which doesn't need to forward the message further. When node G wants to join the group, it sends a join request to D; D already has a forwarding entry for g_x, and will simply add G to the existing entry. When a host sends a packet to a group, its DR router will send it as unicast packet (through a technique called "encapsulation") to the RP router of the group, then the RP router sends the packet as a multicast packet along the tree established. For example, the host s sends a packet to group g_x; its DR router S sends the packet as a unicast packet to the RP router B. B has forwarding entry for g_x and forwards it to D and E, and D in turn forwards it to F and G.

PIM-SM is similar to a server-based solution to some degree. The main difference is that when packets reach the server (the RP router in this case), they are distributed to all members through a tree instead of many unicast connections. PIM-SM also has some other advanced features such as providing support to switch to a source-specific tree (i.e., a multicast tree rooted at the source for a particular source). We won't discuss these features here, and interested readers can refer to the related literature.

4.4. Core-Based Tree (CBT)

All the protocols discussed above build unidirectional multicast trees; at a tree router, multicast packets are expected to arrive from one interface and will be sent out to a list of outgoing interfaces. In PIM-SM, a tree is shared — all source nodes send packets to the RP and the RP sends them over the tree. However, DVMRP and MOSPF use source-specific trees. Thus, if there are multiple sending sources for a group, then multiple trees must be established. For example, if a host at A's subnet sends a packet to group g_x, a new tree rooted at A will be established in MOSPF or DVMRP. The implication is a scalability issue — the more trees, the more forwarding entries a router has to maintain and the more processing overhead for multicast forwarding. Another protocol, Core-Based Tree (CBT) [3,4] builds a single bidirectional shared tree for a group. Although this protocol hasn't been and may never be widely deployed, its idea has been shared in PIM-SM, and a newer interdomain multicast routing protocol called *Border Gateway Multicast Protocol* (BGMP) is based on it.

Similar to PIM-SM, CBT has a core router for a group and uses an explicit join mechanism for tree construction. When a router joins a group, it sends an explicit join request message toward the core until the message reaches the core or a node that is already in the tree, and then a new branch is established. Forwarding entry at a router is for per group (i.e., *g, list_of_interfaces*) (Section 2.3). When a node want to sends a packet to a group, it sends it toward the core until it reaches a node that is already in the tree, and the packet will travel in the multicast tree as a multicast packet from that point. The main difference compared with PIM-SM is that the packet doesn't need to reach the core.

5. INTERDOMAIN MULTICAST ROUTING PROTOCOLS AND ALGORITHMS

Two of the above routing protocols (DVMRP, MOSPF) are for intradomain multicast only. The other two (CBT and PIM-SM) were actually designed to support interdomain multicast as well — they both require only the underlying unicast routing protocol to provide a next hop to a core or RP router and that is readily supported by both intradomain routing protocols and interdomain routing protocols [e.g., Border Gateway Protocol (BGP) [19]]. However, because of some limitations and other emerging problems as Internet multicast grows out of the old experimental multicast backbone (MBone) [2], there is a pressing need to develop new protocols to better support multicast at the interdomain level.

Over the years, several protocols have been developed and considered by IETF to provide scalable hierarchical Internetwide multicast. The first step toward scalable hierarchical multicast routing is Multiprotocol Extensions to BGP4 (MBGP) [5], which extends BGP to carry multiprotocol routes (i.e., besides the "traditional" IP unicast routes). In the MBGP/PIM-SM/MSDP architecture [2], MBGP is used to exchange multicast routes and PIM-SM is used to connect group members across domains, while another protocol, Multicast Source Discovery Protocol (MSDP) [14], was developed to exchange information of active multicast sources among RP routers across domains.

The MBGP/PIM-SM/MSDP architecture has scalability problems and other limitations, and is recognized as

a near-term solution [2]. To develop a better long-term solution, a more recent effort is the MASC/BGMP architecture [24].

5.1. The MASC/BGMP Architecture

In the MASC/BGMP [24,37] architecture, border routers run Border Gateway Multicast Protocol (BGMP) to construct a bidirectional "shared" tree similar to a CBT tree for a multicast group. The shared tree is rooted at a "root domain" (instead of a single core router) that is mainly responsible for the group (e.g., the domain where the group communication initiator resides). To solve the difficult problem of mapping a multicast group to a RP or core router associated with PIM-SM or CBT, BGMP relies on a hierarchical multicast group address allocation protocol called the *Multicast Address-Set Claim Protocol* (MASC) to map a group address to a root domain and an interdomain routing protocol (BGP/MBGP) to carry "group route" information (i.e., how to reach the root domain of a multicast group).

MASC is used by one or more nodes of a MASC domain to acquire address ranges to use in a domain. Within the domain, multicast addresses are uniquely assigned to clients using an intradomain mechanism. MASC domains form a hierarchical structure in which a "child" domain (customer) chooses one or more "parent" (provider) domains to acquire address ranges using MASC. Address ranges used by top-level domains (domains that don't have parents) can be preassigned and can then be obtained by child domains. This is illustrated in Fig. 7, in which A, D, and E are backbone domains, B and C are customers of A, while B and C have their own customers F and G, respectively. A has already acquired address range 224.0.0.0/16 from which B and C obtain address ranges 224.0.128.0/24 and 224.0.1.1/25, respectively.

Using this hierarchical address allocation, multicast "group routes" can be advertised and aggregated much like unicast routes. For example, border router B_1 of domain B advertises *reachability* of root domains for groups in the range of 224.0.128.0/24 to A_3 of domain A, and $A_1(A_4)$ advertises the aggregated 224.0.0.0/16 to $E_1(D_1)$ in domain $E(D)$. Group routes are carried through MBGP and are injected into BGP routing tables of border routers. BGMP then uses such "group routing information" to construct shared multicast trees to distribute multicast packets.

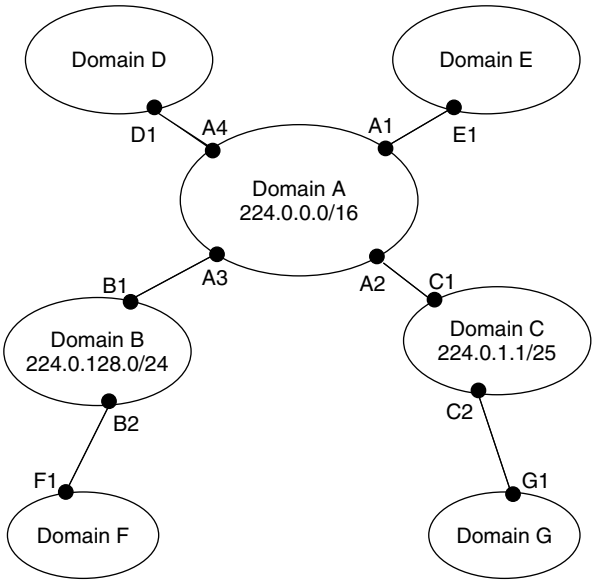


Figure 7. Address allocation using MASC, adopted from Ref. 24.

BGMP constructs a bidirectional shared tree for a group rooted at its root domain through explicit join/prune as in CBT. An example tree is illustrated in Fig. 8. A BGMP router in the tree maintains a *target list* that includes a *parent target* and a list of *child targets*. A parent target is the next-hop BGMP peer toward the root domain of the group. A child target is either a BGMP peer or an MIGP (Multicast Interior Gateway Protocol, i.e., MOSPF or PIM-SM) component of this router from which a join request was received for this group. For example, assume domain B is the root domain, then at node C_2, the parent target is node A_2 and a child target may be an interface of its own that connects to another tree router in its domain. Data packets received for the group will be forwarded to all targets on the list except the one from which the data packet originated. BGMP router peers maintain persistent TCP connections with each to exchange BGMP control messages (join/prune, etc.).

In the BGMP architecture, a source doesn't need to join the group in order to send data. When a BGMP router receives data packets for a group for which it doesn't have a forwarding entry, it will simply forward packets to the next-hop BGMP peer toward the root domain of the group. Eventually they will hit a BGMP router that has a forwarding state for that group or a BGMP router in the root domain. For example, if a node in domain D wants to send a packet to group 224.0.128.5, based on the group address (belonging to domain B), the packet will be sent to node D_1, then to A_4, and then A_3. Node A_3 is already in the tree; thus it will forward the packet over the multicast tree (to its interdomain peer B_1 and interior neighbor A_a). BGMP can also build source-specific branches, but only when needed (i.e., to be compatible with

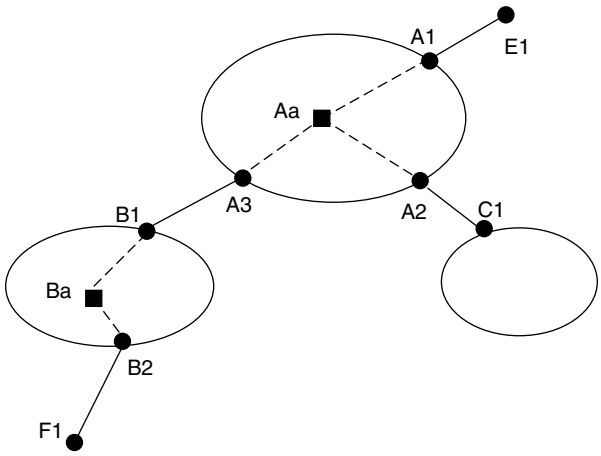


Figure 8. An interdomain multicast tree (solid lines are tree links). Within a domain, an intradomain multicast routing protocol builds an intradomain multicast (dashed lines).

source-specific trees used by some intradomain multicast protocols such as DVMRP and MOSPF), or to construct trees for source-specific groups.

6. CONCLUSIONS

Today, many applications involve group communications. Multicast was conceived as a mechanism to efficiently support such a communication need. However, multicast support at the network level is only rudimentary in traditional telephone networks. Though extensive work on multicast in IP networks has been started since the early 1980s, widespread availability of multicast service in the Internet is still not any time soon. Much research and engineering work is still to be done. This article focused mostly on multicast in IP networks and the Internet, especially routing protocols and algorithms that constitute the core of multicast support.

The multicast routing problem can theoretically formulated as the Steiner tree problem. However, because of the routing environment and some network requirements, traditional Steiner tree algorithms are not readily applicable for multicast routing. Several routing algorithms and protocols have been developed, and some are standardized by the IETF for the Internet. They include DVMRP, MOSPF, CBT, PIM-SM, and the more recent MBGP. In this article, we described a big picture of IP multicast and then gave an overview of those protocols and algorithms. Interested readers can refer to the references cited for more detailed specific information on any of them. This field is still under very active development, and many new advances are being made. Interested readers may refer to the most recent research literature and IETF documents for more recent developments.

BIOGRAPHIES

Aiguo Fei (afei@acm.org) received the B.S. degree in physics in 1995 from Fudan University, Shanghai, China; the M.S. degree in physics and the M.S. and Ph.D. degrees in computer science in 1996, 1998, and 2001, respectively, from University of California, Los Angeles. He joined a startup company in Silicon Valley, California in 2001 as a research engineer. His area of interests are multicast in IP networks, QoS support in next-generation IP networks, network and graph algorithms, and network intrusion detection and statistical anomaly.

Mario Gerla (gerla@cs.ucla.edu) was born in Milan, Italy. He received a graduate degree in engineering from the Politecnico di Milano, in 1966, and the M.S. and Ph.D. degrees in engineering from UCLA in 1970 and 1973, respectively. He joined the Faculty of the UCLA Computer Science Department in 1977. His research interests cover the performance evaluation, design, and control of distributed computer communication systems; high-speed computer networks; wireless LANs (Bluetooth); and ad hoc wireless networks. He has been involved in the design, implementation, and testing of wireless ad hoc network protocols (channel access, clustering, routing, and transport) within the DARPA WAMIS, GloMo projects and most recently the ONR MINUTEMAN project. He has also carried out design and implementation of QoS routing, multicasting protocols, and TCP transport for the next-generation Internet (see www.cs.ucla.edu/NRL for the most recent publications).

BIBLIOGRAPHY

1. Internet Engineering Task Force. *http://www.ietf.org/*.
2. K. Almeroth, The evolution of multicast: From the MBone to inter-domain multicast to Internet2 deployment, *IEEE Network* (Jan./Feb. 2000).
3. A. Ballardie, *Core Based Trees (CBT version 2) Multicast Routing: Protocol Specification*, IETF RFC 2189, Sept. 1997.
4. A. Ballardie, P. Francis, and J. Crowcroft, Core based trees (CBT), *Proc. ACM SIGCOMM'93*, Sept. 1993, pp. 85–95.
5. T. Bates, R. Chandra, D. Katz, and Y. Rekhter, *Multiprotocol Extensions for BGP-4*, IETF RFC 2283, Feb. 1998.
6. B. Cain et al., Internet group management protocol, version 3, *Internet draft: draft-ietf-idmr-igmp-v3-07.txt*, March 2001.
7. Ch.-H. J. Wu, and J. D. Irwin, *Emerging Multimedia Computer Communication Technologies*, Prentice-Hall, 1998.
8. D. E. Comer, *Computer Networks & Internets with Internet Applications*, 3rd ed., Prentice-Hall, 2001.
9. T. H. Cormen, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms*, The MIT Press, 1990.
10. S. Deering, D. Estrin, D. Farinacci et al., Protocol independent multicast-sparse mode (pim-sm): motivation and architecture, *IETF Internet draft: draft-ietf-idmr-pim-arch-05.txt{ps}*, Aug. 1998.
11. S. Deering et al., The pim architecture for wide-area multicast routing, *IEEE/ACM Trans. Network.* **4**(2): 153–162 (April 1996).
12. C. Diot, W. Dabbou, and J. Crowcroft, Multipoint communication: A survey of protocols, functions, and mechanisms, *IEEE J. Select. Areas Commun.* **15**(3): 277–290 (April 1997).
13. D. Estrin et al., *Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification*, IETF RFC 2362, June 1998.
14. D. Farinacci et al., Multicast source discovery protocol (msdp), *IETF Internet draft: draft-ietf-msdp-spec-06.txt*, 2000.
15. W. Fenner, *Internet Group Management Protocol, Version 2*, IETF RFC 2236, 1997.
16. Joe Ganley, *http://ganley.org/steiner/*.
17. M. Goncalves and K. Niles, *IP Multicasting: Concepts and Applications*, McGraw-Hill, 1999.
18. C. Hedrick, *Routing Information Protocol*, IETF RFC 1058, 1988.
19. C. Huitema, *Routing in the Internet*, Prentice-Hall, 1995.
20. F. Hwang, D. Richards, and P. Winter, *The Steiner Tree Problem*, Elsevier, 1992.
21. D. Jungnickel, *Graphs, Networks and Algorithms*, Algorithms and Computation in Mathematics, Springer, 1999.
22. S. Keshav, *An Engineering Approach to Computer Networking: ATM Networks, the Internet, and the Telephone Network*, Addison-Wesley, 1997.

23. D. Kosiur, *IP Multicasting*, Wiley, 1998.
24. S. Kumar et al., The MASC/BGMP architecture for inter-domain multicast routing, *Proc. ACM SIGCOMM'98*, Sept. 1998, pp. 93–104.
25. U. Manber, *Introduction to Algorithms: A Creative Approach*, Addison-Wesley, 1989.
26. C. K. Miller, *Multicast Networking and Applications*, Addison-Wesley, 1999.
27. J. Moy, Multicast routing extensions for ospf, *Commun. ACM* **37**: 61–66 (Aug. 1994).
28. J. Moy, *Multicast Routing Extensions to OSPF*, RFC 1584, March 1994.
29. J. Moy, *Ospf Version 2*, IETF RFC 2328, April 1998.
30. C. Partridge, D. Waitzman, and S. Deering, *Distance Vector Multicast Routing Protocol*, RFC 1075, 1988.
31. R. Perlman, *Interconnections: Bridges, Routers, Switches, and Internetworking Protocols*, 2nd ed., Addison-Wesley, 1999.
32. L. L. Peterson and B. S. Davie, *Computer Networks: A Systems Approach*, 2nd ed., Morgan Kaufmann, 1999.
33. S. S. Skiena, *The Algorithm Design Manual*, Springer, 1997.
34. H. Takahashi and A. Matsuyama, An approximate solution for the Steiner problem in graphs, *Math. Jpn.* **24**: 573–577 (1980).
35. A. S. Tanenbaum, *Computer Networks*, 3rd ed., Prentice-Hall, 1996.
36. R. E. Tarjan, *Data Structures and Network Algorithms*, Society for Industrial and Applied Mathematics, 1983.
37. D. Thaler, D. Estrin, and D. Meyer, Border gateway multicast protocol (BGMP): Protocol specification, *IETF Internet draft: draft-ietf-bgmp-spec-02.txt*, Nov. 2000.
38. R. Wittmann and M. Zitterbart, *Multicast Communication: Protocols and Applications*, Morgan Kaufmann, Academic Press, San Francisco, 2001.

MULTIDIMENSIONAL CODES

JOHN M. SHEA
TAN F. WONG
University of Florida
Gainesville, Florida

1. INTRODUCTION

The term *multidimensional codes* is used in several different contexts relating to modern communications. For instance, trellis coding with multidimensional modulation [1–3] is sometimes referred to as *multidimensional coding*. Trellis coding with multidimensional modulation uses multiple modulation symbols that map to a symbol of greater than two dimensions. Certain algebraic geometry codes that are defined using projective algebraic curves over Galois fields [4] are also referred to as *multidimensional codes*. Other types of codes that are multidimensional in nature are the two-dimensional burst identification codes of Abdel-Ghaffar et al. [5] and the two-dimensional dot codes of van Gils [6].

In this article, we consider multidimensional codes in which the bits are encoded by a series of orthogonal parity checks that can be represented as coding in different dimensions of a multidimensional array. In Section 2, we provide a brief introduction to product codes and some of their properties. In Section 3, we discuss the properties of product codes constructed from single parity-check codes, and in Section 4 we discuss soft-decision decoding of these codes. Finally, in Sections 5–7, we present a class of multidimensional codes and provide performance results for two applications of these codes.

2. PRODUCT CODES

Product codes were introduced by Elias in 1954 [7] as a way to develop a code that could achieve vanishingly small error probability at a positive code rate. The scheme proposed by Elias uses an iterative coding and decoding scheme in which each decoder improves the channel error probability for the next decoder. In order to ensure that any errors at the outputs of one decoder appear as independent error events in each codeword input to the next decoder, Elias proposed encoding each information bit using a series of orthogonal parity checks. He termed this technique "iterative" coding. His coding scheme is now commonly referred to as a *product code*. In particular, the coding scheme he proposed is a systematic, multidimensional product code in which the number of dimensions can be chosen to achieve arbitrarily low bit error probability.

Product codes are the most common form of multidimensional code. A product code of dimension p is generated in such a way that each information bit is encoded p times. Product codes are typically formed using linear block codes [8]. Suppose that we have p (not necessarily different) block codes $C_1, C_2, \ldots, C_p$ with blocklength $n_1, n_2, \ldots, n_p$ and information length $k_1, k_2, \ldots, k_p$. Then the p-dimensional product of these codes is a block code C, with blocklength $n = n_1 n_2 \cdots n_p$ and information length $k_1 k_2 \cdots k_p$. The constituent codes $\{C_i\}$ are said to be subcodes of C [9].

For a two-dimensional product code, the code can be visualized as a rectangular array in which each column is a codeword in C_1 and each row is a codeword in C_2. Let n_i denote the blocklength of code C_i, and let k_i denote the number of information bits conveyed by each codeword of code C_i. We say that code C_i is a (n_i, k_i)-block code. Suppose that C_1 and C_2 are systematic codes. A diagram that illustrates the construction of Elias' systematic two-dimensional product code is shown in Fig. 1. One possible way to encode the information is as follows:

1. Place the information bits in the $k_1 \times k_2$ submatrix.
2. For each of the first k_2 columns, calculate $n_1 - k_1$ parity bits using code C_1 and append those to that column.
3. For each of the first k_1 rows, calculate $n_2 - k_2$ parity bits using code C_2 and append those to that row.
4. For each of the last $(n_1 - k_1)$ rows, calculate $(n_2 - k_2)$ parity bits from code C_2 and append those to that row. This last set of parity bits uses the parity bits from code C_1 and encodes them with code C_2, and are thus known as *parity-on-parity* bits.

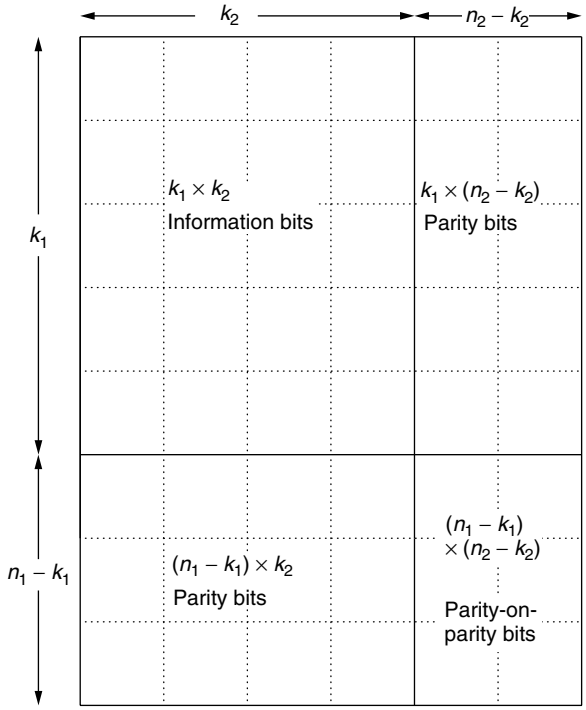


Figure 1. A two-dimensional product code.

Note that in step 4, the parity-on-parity bits have the same values if they are instead constructed using code C_1 on the parity bits from code C_2.

Product codes have many properties that are derived from their subcodes. Some of the most commonly used block codes are the cyclic codes [8]. If $\mathbf{v}$ is a codeword of a cyclic code C, then any cyclic shift of $\mathbf{v}$ is also a codeword of C. Peterson and Weldon [10] proved the following theorem for two-dimensional product codes constructed from cyclic subcodes.

Theorem 1. Suppose that C_1 and C_2 are cyclic codes with length n_1 and n_2, where n_1 and n_2 are relatively prime. Let $i_1 \equiv i \bmod n_1$, and let $i_2 \equiv i \bmod n_2$, for $i = 0, 1, 2, \ldots, n_1 n_2 - 1$. Then the product of C_1 and C_2 is a cyclic code if the codeword $\mathbf{v} = (v_0, v_1, \ldots, v_{n_1 n_2 - 1})$ is constructed such that the symbol v_i is the symbol in the (i_1, i_2)th position of the rectangular array representation.

The mapping from i to (i_1, i_2) results in a cyclic enumeration of i that wraps around the edges of the $n_1 \times n_2$ rectangular array. For example, consider $n_1 = 5$ and $n_2 = 3$. Then the positions of i in the rectangular array are as shown in Fig. 2a. Note that a right cyclic shift of the codeword $\mathbf{v}$ corresponds to a right cyclic shift and downward cyclic shift of the rectangular matrix, as is illustrated in Fig. 2b. Thus, a cyclic shift of the codeword $\mathbf{v}$ results in another valid codeword.

Products of more than two codes are also cyclic codes under similar constructions (see Corollary II of Ref. 9). Furthermore, the generator polynomial [8] for the product code is shown to be a simple function of the generator polynomials for the subcodes. Let $g_i(X)$ be the generator polynomial for the ith subcode, and let $g(X)$ be the

(**a**)

0	10	5
6	1	11
12	7	2
3	13	8
9	4	14

(**b**)

14	9	4
5	0	10
11	6	1
2	12	7
8	3	13

Figure 2. Bit ordering to convert 5×3 product code to a cyclic code: (**a**) original order; (**b**) order after right cyclic shift.

generator polynomial for the product code. The generator polynomial for the two-dimensional product code is given by (see Theorem III of Ref. 9)

$$g(X) = \text{GCD}\ \{g_1(X^{bn_2}) g_2(X^{an_1}), X^{n_1 n_2} - 1\}$$

where $\text{GCD}(y, z)$ is the greatest common divisor of y and z, and a and b are integers satisfying $an_1 + bn_2 \equiv 1 \bmod n_1 n_2$. For example, for the 5×3 code of Fig. 2, $a = 2$ and $b = 2$ will satisfy the equation $(2)(5) + (2)(3) \equiv 1 \bmod 15$. Note that these values come from the structure of the cyclic form of the product code. This is visible in Fig. 2, in which the separation $(\bmod n_1 n_2)$ between neighboring positions is $an_1 = 10$ in any row and $bn_2 = 6$ in any column.

Let $d_1, d_2, \ldots, d_p$ denote the minimum distances of subcodes $C_1, C_2, \ldots, C_p$, respectively. Elias [7], shows that the minimum distance d for the product code C is the product of the minimum distances of the subcodes, $d = d_1 d_2 \cdots d_p$. Thus, product codes offer a simple way to construct a code with a large minimum distance from a set of shorter codes with smaller minimum distances.

We present some results on the error correction capability of product codes that are based on the structure of the codes. We note that these results are not necessarily achievable with most hard-decision decoding algorithms. The random-error-correction capability, which is the maximum number of errors that a code is guaranteed to correct, is thus given by

$$t = \left\lfloor \frac{d-1}{2} \right\rfloor = \left\lfloor \frac{\left(\prod_{i=1}^{p} d_i \right) - 1}{2} \right\rfloor.$$

Some error-control codes are able to correct more than t errors if the errors occur in bursts. A single error burst of length B_p occurs if all the errors in the codeword are constrained to B_p consecutive symbols of the codeword. Cyclic product codes are particularly useful for burst error correction. Again, consider a two-dimensional cyclic product code. Let t_1 and t_2 denote the random error

correction capability of subcodes C_1 and C_2, respectively. Let B_1 and B_2 denote the maximum length of an error burst that is guaranteed to be corrected by subcodes C_1 and C_2, respectively. Then code C_i can correct all errors that are constrained to B_i consecutive positions, regardless of the weight of the error event. The value B_i is said to be the burst error correction capability of subcode C_i. Let B_p denote the burst error correction capability of the product code. Then it is shown in [9] that B_p satisfies the following bounds:

$$B_p \geq n_1 t_2 + B_1$$

and

$$B_p \geq n_2 t_1 + B_2.$$

Several researchers have investigated the burst error correction capability of product codes constructed from single parity-check codes. This research is discussed in the following section.

3. PRODUCTS OF SINGLE PARITY-CHECK CODES

A particular product code that has drawn considerable attention is the p-time product of single parity-check (SPC) codes. We will refer to these codes as product SPC codes. Single parity-check codes are $(k+1, k)$ codes for which one parity bit is added for each k input bits. Typically, the parity bit is computed using *even parity*, in which case the sum of all the bits in the codeword is an even number. The parity-check code is a cyclic code with generator polynomial $g(X) = X + 1$. Product-SPC codes appear in the literature at about the same time from two different groups of authors, Calabi and Haefeli [11] and Gilbert [12]. Interestingly, each of these authors further attributes the original code idea to other researchers. Calabi and Haefeli attribute the code to C. Hobbs of the Communications Laboratory, Air Force Cambridge Research Center. Gilbert attributes a two-dimensional form of the code to W. D. Lewis. Apparently, the original contributions by W. D. Lewis were unpublished, but some extensions were patented as U.S. patents 2,954,432 and 2,954,433. Product codes that use SPC codes as subcodes have been called Hobbs' codes [11] or Gilbert codes [13,14].

Let C_i be a single parity-check code of length n_i for $1 \leq i \leq p$, and let $n = n_1 n_2 \cdots n_p$. Then the p-dimensional product of C_i, $1 \leq i \leq p$, is denoted by C and has blocklength n. Note that each information bit participates in exactly one parity-check equation in each dimension. The number of parity-check equations in which a bit participates is referred to as its *density* [13,15]. Since each bit participates in exactly p checks, the parity-check matrix can be constructed to have exactly p ones in every column. Such a parity-check matrix is known as a *regular* low-density parity-check matrix [15]. Suppose that the 5×3 product code shown in Fig. 2 is constructed from even SPC codes. Let v_{ij} denote the (i,j)th code symbol in the rectangular array representation. Then the code symbols must satisfy the row parity-check equations

$$\sum_{j=0}^{2} v_{ij} = 0, i = 0, 1, 2, 3, 4$$

and the column parity-check equations

$$\sum_{i=0}^{4} v_{ij} = 0, j = 0, 1, 2$$

Thus, if the codeword $\mathbf{v}$ is constructed using the bit ordering described in Theorem 1, a low-density representation for the parity-check matrix $\mathbf{H}$ is given by

$$\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1
\end{bmatrix}.$$

This low-density parity-check matrix can be used in implementing a soft-decision decoder for the code.

The SPC code is a cyclic code, so if $n_1, n_2, \ldots, n_p$ are relatively prime, then the product code will be cyclic. For the two-dimensional code with n_1 and n_2 relatively prime, the generator polynomial is given by [13]

$$\begin{aligned} g(X) &= \mathrm{LCM}\{X_1^n + 1, X_2^n + 1\} \\ &= \frac{(X^{n_1}+1)(X^{n_2}+1)}{X+1} \end{aligned}$$

where $\mathrm{LCM}(y, z)$ denotes the least common multiple of y and z. For example, for the 5×3 product SPC code, the generator polynomial is given by

$$\begin{aligned} g(X) &= \frac{(X^5+1)(X^3+1)}{X+1} \\ &= X^7 + X^6 + X^5 + X^2 + X + 1 \end{aligned}$$

Note that codes formed this way are not only cyclic, but are *palindromic* [13], in which the code is the same if each codeword is read backward. Thus, for the 5×3 code, the *reciprocal* [8] of $g(X)$ is

$$\begin{aligned} X^7 g(X^{-1}) &= 1 + X + X^2 + X^5 + X^6 + X^7 \\ &= g(X) \end{aligned}$$

A parity-check polynomial for a cyclic code can be found [8] using the parity-check polynomial $h(X) = (X^n + 1)/g(X)$. However, the parity-check matrix formed using $h(X)$ is seldom a low-density matrix.

The burst error correction capabilities of these codes have been investigated [11–14,16,17]. Burst error detection capabilities were also investigated [13,14]. We present a summary of some of the most important results here. Neumann [13] points out that some of the product SPC codes are "fire" codes [8], which are another class of block codes designed for burst error correction. We consider the maximum-length error burst that the code can correct, and we denote the length of such a burst by B_p. Consider first the case of a two-dimensional $n_1 \times n_2$ product code,

where n_1 and n_2 are relatively prime. Let π_i denote the smallest prime divisor of n_i, and define

$$b_i = \left(\frac{\pi_i - 1}{\pi_i}\right) n_i$$

Then the code can correct all single error bursts up to length $B_p = \min\{b_1, b_2, \lfloor (n_1 + n_2 + 2)/3 \rfloor\}$. Thus for the 5×3 product SPC code, the single-burst error correction capability is

$$\begin{aligned} B_p &= \min\left\{\frac{5-1}{5} \cdot 5, \frac{3-1}{3} \cdot 3, \left\lfloor \frac{3+5+2}{3} \right\rfloor\right\} \\ &= \min\{4, 2, 3\} \\ &= 2 \end{aligned}$$

A *solid* burst error is one in which every bit in the burst is received in error. Then it is shown [14] that the two-dimensional product SPC code can correct all solid burst errors of length $\min\{n_1, n_2\} - 1$.

If a two-dimensional cyclic product SPC code is used for single-burst error detection, then its error detection capability is easily derived from the properties of cyclic codes [8,10]. Any burst of length $n - k$ can be detected for an (n, k) cyclic code. Thus, any burst of length up to $n_1 + n_2 - 1$ can be detected for a $n_1 \times n_2$ cyclic product SPC code. Neumann also shows that the code has the capability to simultaneously correct B_p errors while detecting burst errors of length almost equal to $\max\{n_1, n_2\}$.

4. DECODING OF PRODUCT SPC CODES

Cyclic product codes can be decoded using a variety of hard-decision and soft-decision decoding algorithms. One advantage of product-SPC codes is that they have very simple and efficient soft-decision decoding algorithms that we discuss in this section. However, we first provide some references to hard-decision decoding algorithms for these codes, particularly for application to burst error correction. Neumann [13] presents a decoding algorithm for single- and double-burst error correction. Bahl and Chien present a threshold decoding algorithm for product SPC codes with multiple error bursts [16] and a syndrome-based decoding algorithm that provides better performance [17].

Several soft-decision decoding algorithms exist for block codes [18–20]. Product SPC codes can be decoded in different ways, but we focus on an iterative decoding process that uses optimal maximum a posteriori probability decoders on each subcode in each dimension. The resulting iterative decoding algorithm is typically not an optimal decoding algorithm but is usually significantly simpler than an optimal decoder. In order to understand the operation of this iterative decoding algorithm, we first focus on the optimal symbol-by-symbol maximum-likelihood decoding algorithm for binary linear block codes.

The useful notion of a *replica* was first introduced in the context of soft-decision decoding by Battail et al. [18]. Consider a codeword $\mathbf{v} = (v_0, v_1, \ldots, v_{n-1})$. A replica for bit v_i is information about bit v_i that can be derived from the code symbols other than v_i. Consider the (7,4) Hamming code. The parity-check matrix for this code can be written as

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}.$$

Let a codeword $\mathbf{v} = (v_0, v_1, \ldots, v_{n-1})$. Then the parity-check equations that involve the first code symbol v_0 are

$$\begin{aligned} v_0 + v_2 + v_3 + v_4 &= 0 \\ v_0 + v_1 + v_2 + v_5 &= 0 \\ v_0 + v_3 + v_5 + v_6 &= 0 \\ v_0 + v_1 + v_4 + v_6 &= 0 \end{aligned}$$

where all sums are modulo 2. Note that these equations are linearly dependent, so not every one conveys unique information. Suppose that we use the first three equations, which are linearly independent. Then the first symbol can be written in terms of the other six symbols in either of three ways:

$$\begin{aligned} v_0 &= v_2 + v_3 + v_4 \\ v_0 &= v_1 + v_2 + v_5 \\ v_0 &= v_3 + v_5 + v_6 \end{aligned}$$

These three equations provide three *algebraic replicas* [18] for v_0 in terms of the other symbols in the code. Note that the parity-check matrix $\mathbf{H}$ is the generator matrix for the dual code [8]. Thus, the replicas can be defined using codewords of the dual code [18]. Suppose that the codeword $\mathbf{v}$ is transmitted using binary phase shift keying (BPSK). In the absence of noise, the symbols at the output of the demodulator can be represented by a vector $\mathbf{x} = (x_0, x_1, \ldots, x_{n-1})$, where x_i represents the binary symbol v_i. When noise is present, we denote the demodulator outputs by $\mathbf{y} = (y_0, y_1, \ldots, y_{n-1})$. Then, in terms of the demodulator outputs, there are three *received replicas* of v_0 in addition to y_0, which may be considered a trivial received replica. For example, for the (7,4) Hamming code described above, y_2, y_3, and y_4 provide one received replica of v_0.

With a random information sequence, let V_i be a random variable denoting the ith code symbol of a codeword $\mathbf{V} = (V_0, V_1, \ldots, V_{n-1})$. The codeword $\mathbf{V}$ is transmitted using BPSK over a memoryless channel. Let $\mathbf{X}$ be a vector that consists of the outputs of the demodulator in the absence of noise. For convenience, let $X_i = +1$ if $V_i = 0$, and let $X_i = -1$ if $V_i = 1$. The loglikelihood ratio (LLR) for X_i is defined by

$$L(X_i) = \log \frac{P(X_i = +1)}{P(X_i = -1)}$$

where the natural (base e) logarithm is used. This term is called the *a priori loglikelihood* ratio. In the presence of noise, the received sequence consists of demodulator outputs $\mathbf{Y} = (Y_0, Y_1, \ldots, Y_{n-1})$. The conditional loglikelihood

ratio for X_i given $Y_i = y_i$ is given by

$$\begin{aligned} L(X_i \mid y_i) &= \log \frac{P(X_i = +1 \mid Y_i = y_i)}{P(X_i = -1 \mid Y_i = y_i)} \\ &= \log \frac{P(Y_i = y_i \mid X_i = +1)}{P(Y_i = y_i \mid X_i = -1)} + \log \frac{P(X_i = +1)}{P(X_i = -1)} \\ &= L(y_i \mid X_i) + L(X_i). \end{aligned}$$

The LLRs are often referred to as "soft" values. The sign of the LLR corresponds to a hard-decision value, while the magnitude of the LLR corresponds to the reliability of the decision.

A symbolwise maximum a posteriori (MAP) probability decoder makes decision on X_i based on the larger of $P(X_i = +1 \mid \mathbf{Y} = \mathbf{y})$ and $P(X_i = -1 \mid \mathbf{Y} = \mathbf{y})$, or, equivalently, the sign of

$$L(X_i \mid \mathbf{Y} = \mathbf{y}) = \log \frac{P(X_i = +1 \mid \mathbf{Y} = \mathbf{y})}{P(X_i = -1 \mid \mathbf{Y} = \mathbf{y})}.$$

Note that in general, the a posteriori loglikelihood ratio for X_i depends not only on Y_i but also other symbols in $\mathbf{Y}$. This dependence corresponds to the other received replicas of X_i. Hence, we need to calculate the contribution of a replica of X_i to the above-mentioned a posteriori LLR. We follow the approach described by Hagenauer et al. [19]. Define $\oplus$ as the addition operator over Galois Field (2) with symbols $+1$ and -1, where $+1$ is the null element (since $+1$ corresponds to 0 in the original binary representation). Suppose that X_j and X_k form a replica of X_i via the relation

$$X_i = X_j \oplus X_k$$

We wish to find the loglikelihood ratio for the replica $X_j \oplus X_k$. By using

$$P(X_j = +1) = \frac{e^{L(X_j)}}{1 + e^{L(X_j)}}$$

with the relationship [19]

$$\begin{aligned} P(X_j \oplus X_k = +1) &= P(X_j = +1)P(X_k = +1) \\ &\quad + [1 - P(X_j = +1)][1 - P(X_k = +1)] \end{aligned}$$

we can write

$$P(X_j \oplus X_k = +1) = \frac{1 + e^{L(X_j)}e^{L(X_k)}}{(1 + e^{L(X_j)})(1 + e^{L(X_k)})}$$

Then, using $P(X_j \oplus X_k = -1) = 1 - P(X_j \oplus X_k = +1)$, the loglikelihood ratio for $X_j \oplus X_k$ can be written as

$$L(X_j \oplus X_k) = \log \frac{1 + e^{L(X_j)}e^{L(X_k)}}{e^{L(X_j)} + e^{L(X_k)}}$$

or equivalently

$$\begin{aligned} L(X_j \oplus X_k) = \log &\frac{[e^{L(X_j)} + 1][e^{L(X_k)} + 1] + [e^{L(X_j)} - 1]}{[e^{L(X_j)} + 1][e^{L(X_k)} + 1] - [e^{L(X_j)} - 1]} \\ &\times \frac{[e^{L(X_k)} - 1]}{[e^{L(X_k)} - 1]} \end{aligned}$$

Note that using the relationship $\tanh(x/2) = (e^x - 1)/(e^x + 1)$, this expression can be simplified to [18,19]

$$\begin{aligned} L(X_j \oplus X_k) &= \log \frac{1 + \tanh(L(X_j)/2)\tanh(L(X_k)/2)}{1 - \tanh(L(X_j)/2)\tanh(L(X_k)/2)} \\ &= 2\ \text{atanh}[\tanh(L(X_j)/2)\tanh(L(X_k)/2)] \end{aligned}$$

In general, if a replica of X_i is given by $X_{j_1} \oplus X_{j_2} \oplus \cdots \oplus X_{j_J}$, then the loglikelihood ratio for the replica is given by

$$2\ \text{atanh}\left[\prod_{k=1}^{J} \tanh \frac{L(X_{j_k})}{2}\right]$$

Note that for high signal-to-noise ratios, this can be approximated by

$$\left[\prod_{k=1}^{j} \text{sgn}(L(X_{j_k}))\right] \cdot \min_{k=1,\ldots,J} |L(X_{j_k})|$$

Thus, the reliability of a replica is generally determined by the smallest reliability of the symbols that make up that replica. More commonly, we wish to determine the conditional loglikelihood ratio for a replica of X_j given the received symbols $y_{j_1}, y_{j_2}, \ldots, y_{j_J}$. Then this conditional LLR can be written as

$$2\ \text{atanh}\left[\prod_{k=1}^{J} \tanh \frac{L(X_{j_k} \mid y_{j_k})}{2}\right]$$

For the $(k+1, k)$ even SPC code, the parity-check matrix is given by

$$\mathbf{H} = [111 \cdots 1]$$

Thus, there is one nontrivial (algebraic) replica for each code symbol. The algebraic replica for X_i is given by

$$\sum_{j=0, j\neq i}^{n-1} \oplus\ X_j = X_0 \oplus X_1 \oplus \cdots \oplus X_{i-1} \oplus X_{i+1} \oplus \cdots \oplus X_{n-1}$$

Thus, the conditional loglikelihood of this replica for X_i given $\mathbf{Y}$ is

$$2\ \text{atanh}\left[\prod_{k=1, k\neq j}^{n-1} \tanh \frac{L(X_k \mid y_k)}{2}\right]$$

Considering this algebraic replica and the trivial replica, the a posteriori loglikelihood ratio $L(X_i \mid \mathbf{Y} = \mathbf{y})$ for a code symbol X_i can be broken down into

1. The a priori loglikelihood ratio $L(X_i)$
2. The conditional loglikelihood ratio $L(y_i \mid X_i)$ of received symbol y_i given X_i
3. *Extrinsic information* $L_e(X_i)$ that is information derived from the replicas of X_i

For the SPC code, the a priori LLR for X_i is $L(X_i)$, and is set to zero initially. Consider transmission over an additive white Gaussian noise (AWGN) channel with code

rate R_c and bit energy-to-noise density ratio E_b/N_0. Then the LLR for the received symbol y_i given X_i is

$$L(y_i \mid X_i) = \log \frac{\exp[-\sigma^{-2}(y_i - 1)^2]}{\exp[-\sigma^{-2}(y_i + 1)^2]} = L_c \cdot y_i$$

where

$$L_c = \frac{2}{\sigma^2} = 4\frac{R_c E_b}{N_0}$$

The extrinsic information, that is, the conditional LLR of the replica of X_i given $\mathbf{y}$, is given by

$$L_e(X_i) = 2 \operatorname{atanh}\left[\prod_{k=0,k\neq i}^{n-1} \tanh\frac{L(X_k \mid y_k)}{2}\right] \approx \left[\prod_{k=0,k\neq i}^{n-1} \operatorname{sgn}(L(X_k \mid y_k))\right] \cdot \min_{k=0,\ldots,n-1,k\neq i} |L(X_k \mid y_k)|$$

Thus, the a posteriori loglikelihood ratio for X_i is given by

$$L(X_i \mid \mathbf{Y} = \mathbf{y}) = L(X_i) + L_c \cdot y_i + L_e(X_i)$$

The extrinsic information represents indirect information [19] about the symbol X_i from other code symbols. In the context of product codes, due to the orthogonal parity-check construction, the extrinsic information about X_i that is derived from a particular component subcode does not involve any received symbols (other than X_i) of any other subcode that involves X_i. Thus, this extrinsic information represents information that is not directly available to the decoders of the other subcodes. The extrinsic information can be used by other subcodes by exchanging extrinsic information between the subcodes in an iterative fashion. Each decoder treats the extrinsic information generated by other decoders as if it were a priori information in its decoding.

To illustrate this iterative decoding algorithm, we consider a two-dimensional product SPC code that is formed from the product of two identical (4,3) SPC codes. Conforming to the usual terminology in the iterative decoding literature [19], we refer to the conditional LLR of X_i given y_i, $L(X_i \mid y_i)$, as the *soft input* to a decoder of a SPC subcode. As discussed previously, this soft input is simply the sum of the *a priori* LLR of X_i, $L(X_i)$, and the conditional LLR of the received symbol y_i given X_i, $L_c y_i$. The decoder generates the extrinsic information of X_i, $L_e(X_i)$, based on the soft inputs as described above. This extrinsic information is then used as the a priori LLR X_i for the decoding of the other component subcode. Extrinsic information may be calculated for only the systematic symbols [19] or for all of the code symbols [21], which may offer some improvement in performance. For the results below, we calculate extrinsic information for all the code symbols. The decoding process alternates between the decoders for the two subcodes until a stopping criterion is met. Then the decoder outputs the a posteriori LLR of X_i, which is simply the sum of the soft input and the extrinsic information from each decoder for X_i. A hard decision on X_i is made based on the sign of this *soft output*.

For example, consider the example illustrated in Fig. 3. The transmitted symbols are shown in Fig. 3a. The rightmost column and the bottom row contain the parity-check symbols, and the other symbols carry information. The LLRs of the received symbols are given in Fig. 3b. We see that the decoder would make five errors (indicated by the shaded symbols) if decisions were made directly using these received values. Figure 3c shows the decoding results of the a posteriori decoding algorithm on the SPC defined along the rows. The number inside the upper triangle for each symbol denotes the soft input for that symbol, while the number inside the lower triangle is the extrinsic information generated by the decoder. Initially, we assume that the a priori LLRs of all the symbols are zero. For instance, the soft input for the symbol in the upper left corner is given by $2 + 0 = 2$. To obtain the soft output for a symbol, we simply need to add the values inside the upper and lower triangles corresponding to that symbol. For instance, the soft output of the symbol in the upper left corner is $2.0 - 0.5 = 1.5$. We see that three errors would result if the decoder were to make decisions based on the soft output after this decoder iteration. The decoding process continues for the SPC defined along the columns. The results are shown in Fig. 3d. We note that the soft input of a symbol is now obtained by adding the received LLR of that symbol (from Fig. 3b) to the extrinsic information generated in the previous decoding process (Fig. 3c). For instance, we obtain the soft input of the symbol in the upper left corner as $2.0 - 0.5 = 1.5$. We see that two errors would result if hard decisions were made at this time. The decoding process then returns to the decoder for the SPC along the rows (Fig. 3e) and then the SPC along the columns (Fig. 3e). We see that all five errors that were initially present are corrected by this iterative decoding process. It is also easy to see that any further decoder iterations will not result in any changes in the hard decisions. So for this example, the decoding process converges. Interested readers are referred to the article by Rankin and Gulliver [21] for additional discussion of the performance of iterative decoding with multidimensional product SPC codes.

5. MULTIDIMENSIONAL PARITY-CHECK CODES

In this section, we define a class of multidimensional parity-check (MDPC) codes [22]. These codes are punctured versions of the M-dimensional product SPC codes discussed above. In particular, the M-dimensional product SPC codes have code rates that decrease as M is increased. By puncturing the majority of the parity-check bits for $M > 2$, the MDPC codes have code rates that increase as M increases.

The parity bits for the MDPC code are determined by placing the information bits into a multidimensional array of size M, where $M > 1$. For a particular value of M, we refer to the M-dimensional parity-check code as an M-DPC code. For the special case of $M = 2$, this code is also referred to as a *rectangular parity-check code* (RPCC). In all that follows, we assume that the size of the array is the same

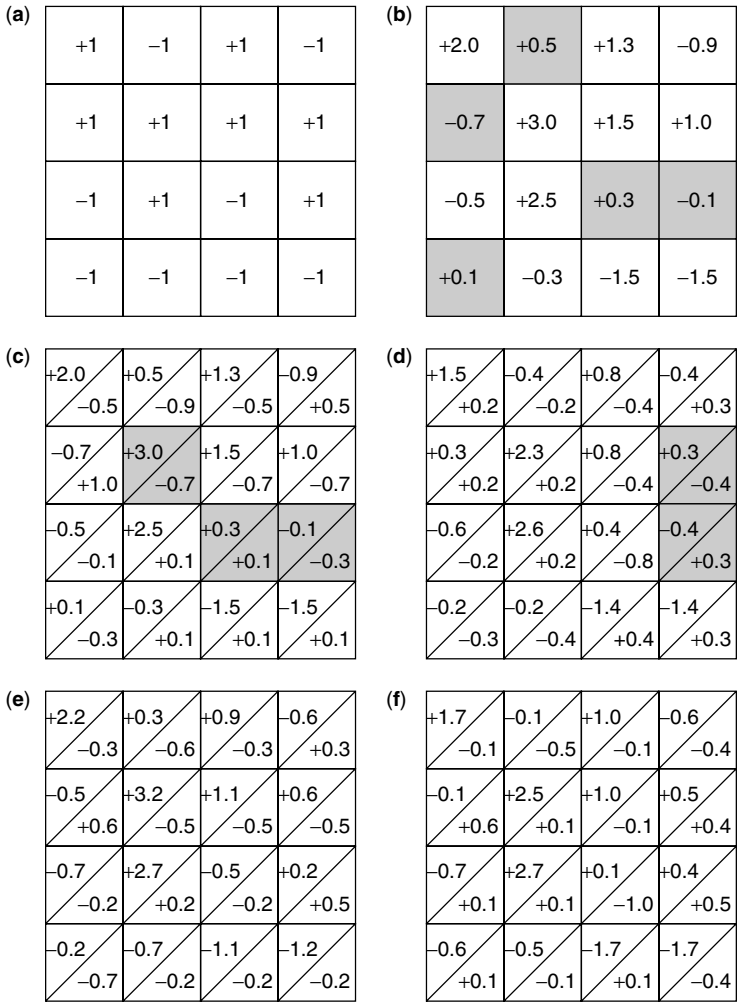


Figure 3. Iterative decoding example of a two-dimensional product SPC: (**a**) transmitted symbols; (**b**) LLRs of received symbols; (**c**) First iteration horizontal SISO decoding; (**d**) First iteration vertical SISO decoding; (**e**) Second iteration horizontal SISO decoding; (**f**) Second iteration vertical SISO decoding.

in each dimension. This is not a requirement in general, but it does minimize the redundancy (and hence the rate penalty) for a given block size and number of dimensions. Suppose that N is the blocklength (the number of input bits that are input to the code to create a codeword), where $N = D^M$. Then D is the size of the M-dimensional array in each dimension. Each M-DPC code is a systematic code in which a codeword consists of D^M information bits and MD parity bits. Then D parity bits are computed for each dimension, where the parity bits can be constructed as the even parity over each of the D hyperplanes of size D^{M-1} that are indexed by that dimension. This is illustrated in Fig. 4 for $M = 2$ and $M = 3$.

More formally, let $u_{i_1,i_2,\ldots,i_M}$ be a block of data bits indexed by the set of M indices $i_1, i_2, \ldots, i_M$. The data bits are arranged in the lattice points of an M-dimensional hypercube of side D. Then the MD parity bits satisfy

$$p_{m,j} = \sum_{i_1} \cdots \sum_{i_{m-1}} \sum_{i_{m+1}} \cdots \sum_{i_M} u_{i_1,\ldots,i_{m-1},j,i_{m+1}\ldots,i_M}$$

for $m = 1, 2, \ldots, M$ and $j = 1, 2, \ldots, D$. Each sum above ranges over D elements, and modulo-2 addition is assumed. Since the M-DPC code produces MD parity bits, the code rate is $D^M/(D^M + MD)$, or equivalently, $(1 + MD^{1-M})^{-1}$. Clearly, as D is increased, the rate of the code becomes very high. For most values of N and M that are of interest, the rate of the M-DPC code increases as M is increased. The minimum code weight is $\min(M + 1, 4)$.

6. BURST ERROR CORRECTION CAPABILITY OF MDPCS WITH ITERATIVE SOFT-DECISION DECODING

MDPC codes can achieve close-to-capacity performance with a simple iterative decoding algorithm in additive white Gaussian noise as well as bursty channels. The iterative decoding algorithm described in Section 4 can be employed with a slight modification. Since the parity-on-parity bits in the product SPC code are punctured in the MDPC code, we do not update the soft inputs corresponding to the parity bits in the iterative decoding process of the MDPC code. Using iterative decoding and with a minimal amount of added redundancy, MDPC codes are very effective for relatively benign channels with possibly occasional long bursts of errors. For example, a three-dimensional parity-check code with a block size of 60,000 can achieve a bit error rate (BER) of 10^{-5} within 0.5 dB of the capacity limit in an AWGN channel while requiring only 0.2% added redundancy. The same code can get to within 1.25 dB of the capacity limit in a bursty channel.

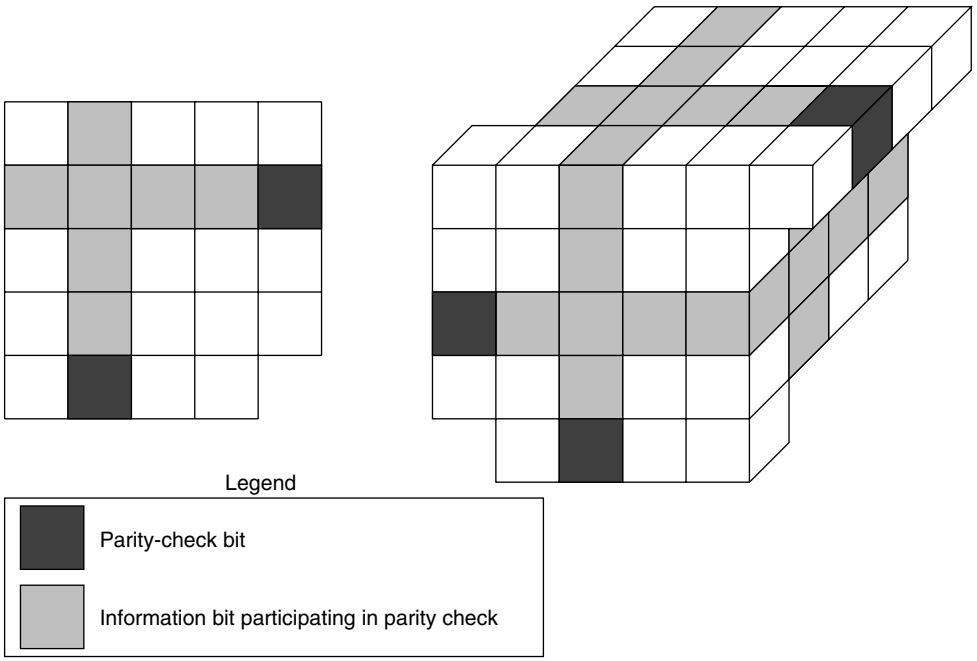


Figure 4. Determination of parity-check bits for $M = 2$ and $M = 3$ MDPC codes.

6.1. Additive White Gaussian Noise Channel

Simulation results of a number of MDPC codes with blocklengths of about 1000, 10,000, and 60,000 data bits over an AWGN channel with BPSK modulation are summarized in Table 1. The results are obtained after 10 iterations for all the codes. However, the decoding process essentially converges after five iterations for all of the MDPC codes that were considered. For instance, the convergence of the iterative decoding process for the 100^2 code is shown in Fig. 5. From Table 1, we conclude that with a block size of 1000 bits, the MDPC codes can achieve a BER of 10^{-5} within 2 dB of the capacity limit.[1] When the block size increases to 10,000 bits, the performance of the MDPC codes is within 1 dB of the capacity. These results are comparable to the ones reported in the article by Chen and McEliece [24], in which codes based on pseudorandom bipartite graphs obtained from computer searches are employed. In comparison, the MDPC codes considered here have much more regular structures, faster convergence rates, and a simpler decoding algorithm. With a block size of approximately 60,000 bits, the 3DPC code 39^3 can achieve a BER of 10^{-5} at 7.9 dB, which is 0.5 dB higher than the capacity limit. Moreover, significant coding gains over uncoded BPSK systems are achieved with very small percentages of added redundancy. For example, using the 21^3 code, a coding gain of 2.3 dB is obtained at 10^{-5} with less than 0.7% redundancy. This accounts for 80.4% of the maximum possible coding gain of 3.25 dB that is allowed by the capacity. It appears that the 3DPC codes are most efficient in terms of attaining the highest percentage of the maximum possible coding gain.

Using the union bound technique [25], we can obtain an upper bound on the bit error probability of the MDPC codes with maximum likelihood (ML) decoding as follows:

$$P_b \le \sum_{i=1}^{D^M} \frac{i}{D^M} \sum_{d=i}^{(M+1)i} W_{i,d} Q\left(\sqrt{\frac{2d E_b/N_0}{1 + M/D^{M-1}}}\right) \qquad (1)$$

where $W_{i,d}$ is the number of codewords with information weight i and codeword weight d. Figure 6 shows the union

[1] Symmetric capacity restricted to BPSK [23] is assumed here.

Table 1. Performance of MDPC Codes Over AWGN Channel

Code	Block Size	Code Rate	E_b/N_0 at 10^{-5} BER (dB)	Coding Gain at 10^{-5} BER (% of Possible Coding Gain)	E_b/N_0 at Capacity (dB)
32^2	1024	0.9412	6.3	3.3 (55.0%)	3.7
10^3	1000	0.9709	6.75	2.85 (63.1%)	4.75
4^5	1024	0.9808	7.25	2.35 (63.8%)	5.3
100^2	10,000	0.9804	6.75	2.85 (70.8%)	5.25
21^3	9261	0.9932	7.3	2.3 (80.4%)	6.35
10^4	10,000	0.9960	7.75	1.85 (80.4%)	6.8
245^2	60,025	0.9919	7.2	2.4 (79.4%)	6.2
39^3	59,319	0.9980	7.9	1.7 (89.1%)	7.4
9^5	59,049	0.9992	8.6	1.0 (88.1%)	8.05

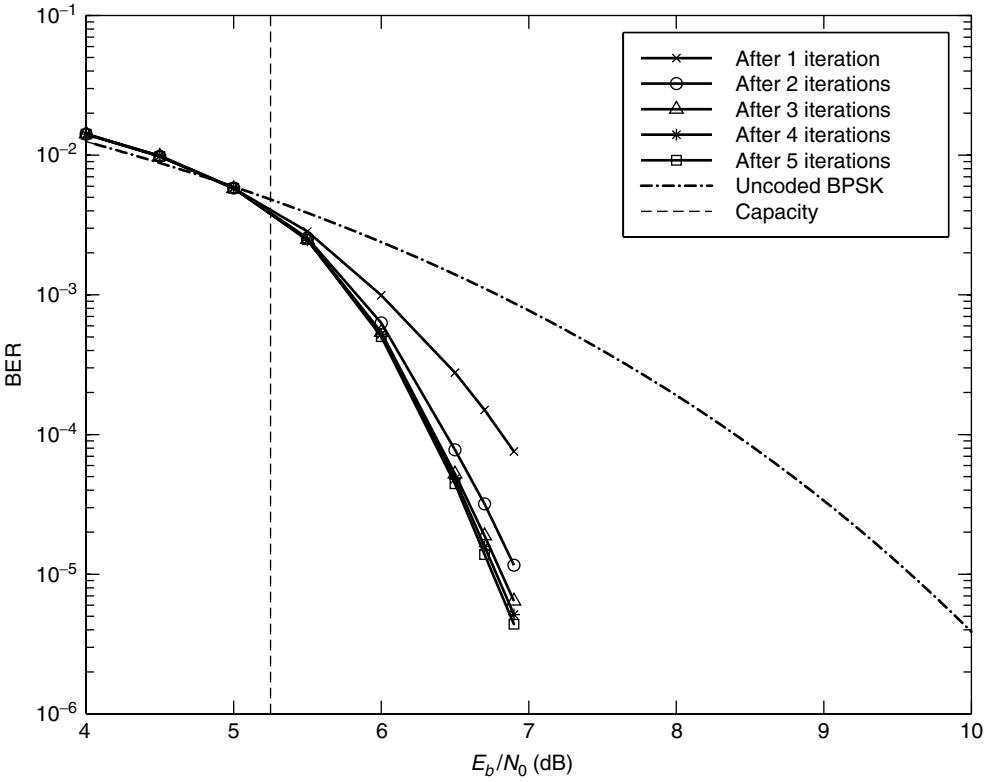


Figure 5. Convergence of iterative decoding process for 100^2 code over AWGN channel.

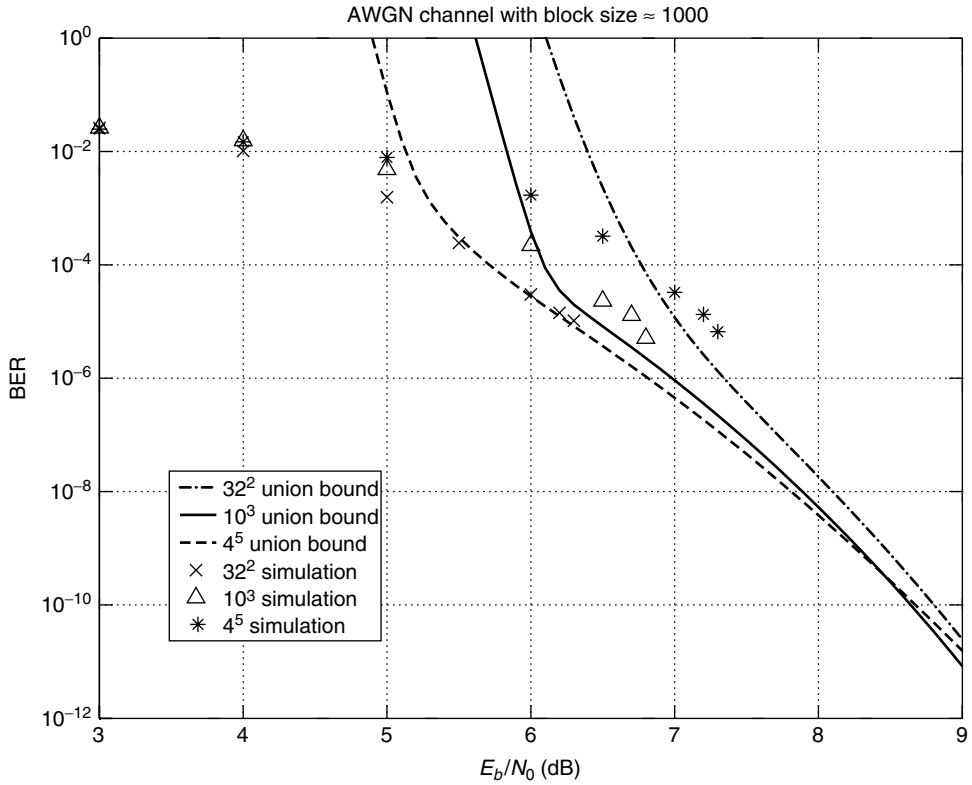


Figure 6. Union bounds and BER performance from simulations of 32^2, 10^3, and 4^5 codes over AWGN channel.

bounds obtained using Eq. (1) for the 32^2, 10^3, and 4^5 codes.[2] Also shown in Fig. 6 are the bit error probabilities of these three codes obtained by the iterative decoder from simulations. We observe from the figure that the BER performance obtained from simulations for the code 32^2 is very close to the corresponding union bound in the high E_b/N_0 region. This indicates that the performance of the iterative decoder is close to that of the ML decoder. For the three- and five-dimensional codes 10^3 and 4^5, the BERs obtained from simulations are poorer than the ones predicted by the respective union bounds. This implies that the iterative decoder becomes less effective as the dimension of the MDPC code increases. Nevertheless, iterative decoding can still provide good coding gains, as shown in Table 1, for MDPC codes with more than two dimensions.

6.2. Bursty Channels

Although the MDPC codes have small minimum distances, they can correct a large number of error patterns of larger weights because of their geometric constructions. With suitable interleaving schemes, the MDPC codes are effective for channels with occasional noise bursts. To examine this claim, we employ the simple two-state hidden Markov model, shown in Fig. 7, to model bursty channels [26]. The system enters state **B** when the channel is having a noise burst. In state **N**, usual AWGN is the only noise. In state **B**, the burst noise is modeled as AWGN with a power spectral density that is B times higher than that of the AWGN in state **N**. It is easy to check that the stationary distribution of the hidden Markov model is $\pi_b = \frac{1-P_n}{1-P_b+1-P_n}$ and $\pi_n = \frac{1-P_b}{1-P_b+1-P_n}$. Assuming that the noise is independent from symbol to symbol after the deinterleaver at the receiver, the conditional LLR $L(y_i \mid X_i)$ is given by

$$L(y_i \mid X_i) = 4R_c\frac{E_b}{N_0}\frac{x}{B} + \log\frac{\exp\left(-\frac{B-1}{B}R_c\frac{E_b}{N_0}(y_i-1)^2\right)+\frac{\pi_b}{\pi_n}}{\exp\left(-\frac{B-1}{B}R_c\frac{E_b}{N_0}(y_i+1)^2\right)+\frac{\pi_b}{\pi_n}} \quad (2)$$

Simulation results of a number of MDPC codes with block sizes of 10,000 and 60,000 bits are summarized in Table 2. In the simulation, $P_b = 0.99$, $P_n = 0.9995$, and $B = 10$ dB. This represents a case where long noise bursts occur occasionally. Random interleavers of sizes equal to the block size of the MDPC codes are employed. The results are obtained after 10 iterations for all the codes. The convergence of the iterative decoding process is similar to the AWGN case. From Table 2, with a block size of 10,000 bits, the 4DPC code 10^4 can achieve a BER of 10^{-5} within 2.4 dB of the capacity limit.[3] When the block size is increased to 60,000 bits, the 5DPC code 9^5 can achieve a BER of 10^{-5} within 1.2 dB of the capacity limit. Although the MDPC codes are not as effective in bursty channels as in AWGN channels, they do provide very significant coding gains with very reasonable complexity as no channel state estimation is needed [26]. In fact, the complicated conditional likelihood ratio calculation in Eq. (2) is not needed since simulation results show that the degradation on the BER performance is very small if the second term on the right-hand side of (2) is neglected.

Using the union bound and assuming that a perfect interleaver is employed so that channel state changes independently from bit to bit, we can obtain the following upper bound on the BER for an ML decoder with perfect channel state information:

$$P_b \le \sum_{i=1}^{D^M}\frac{i}{D^M}\sum_{d=i}^{(M+1)i}W_{i,d}\sum_{k=0}^{d}\binom{d}{k}\pi_b^k\pi_n^{d-k}Q \times\left(\sqrt{\frac{2E_b/N_0}{1+M/D^{M-1}}}\cdot\frac{d-k+k/\sqrt{B}}{\sqrt{d}}\right) \quad (3)$$

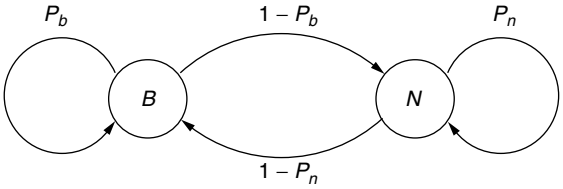


Figure 7. Hidden Markov model for bursty channel.

[2] Here, the weight enumerator coefficients $W_{i,d}$ are obtained approximately by the Monte Carlo method. The first 30 terms in Eq. (1) are used to approximate the union bound.

[3] The capacity here is obtained by averaging the symmetric capacities under the normal and bursty states based on the stationary distribution of the hidden Markov model. This corresponds to the case that a perfect interleaver is employed so that for a given bit, the channel state is independent of the channel states of the other bits, and perfect channel state information is available at the receiver [27].

Table 2. Performance of MDPC Codes Over Bursty Channel

Code	Block Size	Code Rate	E_b/N_0 at 10^{-5} BER (dB)	Coding Gain at 10^{-5} BER (% of Possible Coding Gain)	E_b/N_0 at Capacity (dB)
100^2	10,000	0.9804	13.7	4.15 (29.5%)	8.4
21^3	9261	0.9932	14.8	3.05 (52.5%)	12.0
10^4	10,000	0.9960	15.5	2.35 (57.5%)	13.1
245^2	60,025	0.9919	14.1	3.75 (55.0%)	11.5
39^3	59,319	0.9980	15.45	2.4 (75.0%)	14.2
9^5	59,049	0.9992	16.6	1.25 (75.9%)	15.4

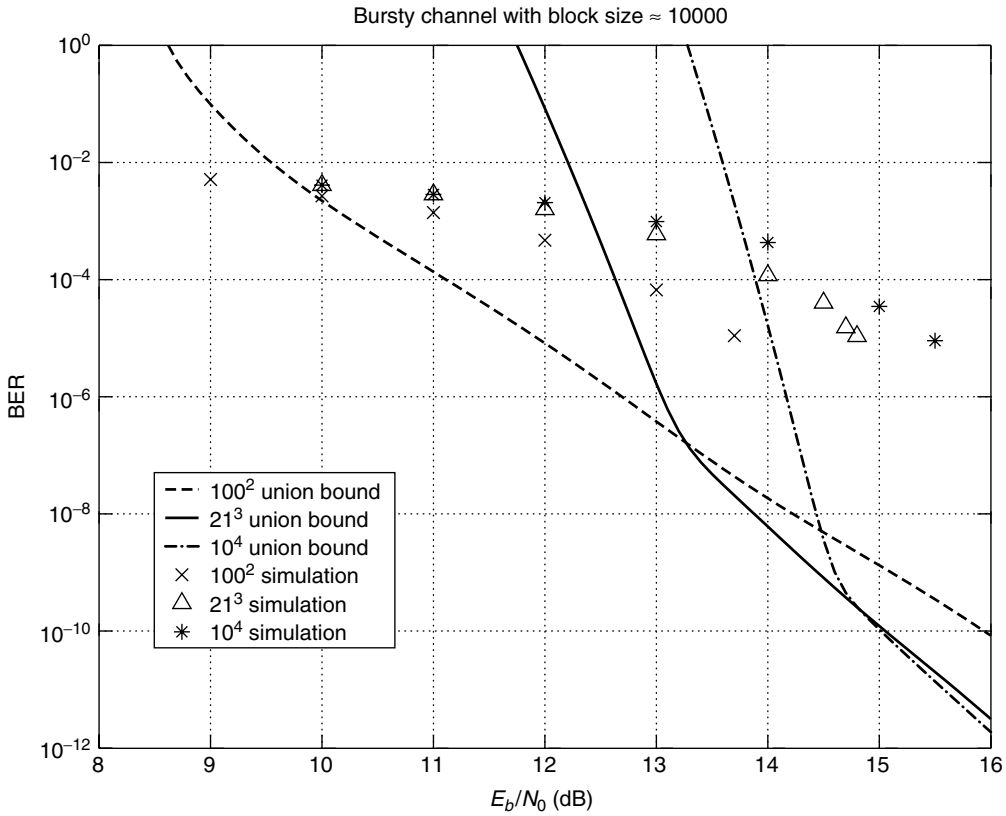


Figure 8. Union bounds and BER performance from simulations of 100^2, 21^3, and 10^4 codes over bursty channel with $P_b = 0.99$, $P_n = 0.9995$, and $B = 10$ dB.

Figure 8 shows the union bounds obtained by this equation for the 100^2, 21^3, and 10^4 codes. Also shown in Fig. 8 are the bit error probabilities of these three codes obtained by the iterative decoder from simulations. We observe from the figure that the BERs obtained from simulations are poorer than those predicted by the union bounds. The reason is threefold. First, the iterative decoder only approximates the MAP decoder. The second, and perhaps the most important, reason is that random interleavers are of the same size as the codewords. These interleavers are not good approximations to the perfect interleaver assumed in the union bound, since such an interleaver would require interleaving across multiple codewords. Third, no channel state information is assumed at the receiver. Nevertheless, the simple iterative decoder and the imperfect interleaver can still give large coding gains as shown in Table 2.

7. CONCATENATED MULTIDIMENSIONAL PARITY-CHECK CODES AND TURBO CODES

Turbo codes [28] are parallel-concatenated convolutional codes that have been shown to provide performance near the capacity limit when very large interleavers (and thus codeword lengths) are used. These codes suffer from an error floor that limits their performance for shorter blocklengths. The error floor is caused by error events that have very low information weight. Thus, these low-weight error-events can be corrected by even a simple outer code. Several authors have investigated the use of an outer code to deal with these low-weight errors. BCH codes have been considered [29–33], and Reed–Solomon codes were also considered [34,35]. However, these codes are typically decoded with algebraic decoders [8] because the complexity is too high for soft-decision decoders for these codes. An alternative approach is to use the multidimensional parity-check codes that are discussed in the previous two sections as outer codes with a Turbo inner code [36,37]. The MDPC codes have simple soft-decision decoders and typically have very high rates that result in less degradation to the performance of the turbo code than most of the other outer codes that have been used. Furthermore, the MDPC codes are good at correcting bursts of errors, such as those that occur at the output of a Turbo decoder.

The results in Fig. 9 illustrate the potential of these codes in regard to improving the error floor. The Turbo code used by itself and in concatenation with the multidimensional parity-check codes is the rate-$\frac{1}{3}$ turbo code with the constituent codes specified in the standards for the cdma2000 and WCDMA third-generation cellular systems [38,39]. The results indicate that the concatenated multidimensional parity-check codes can reduce the error floor by many orders of magnitude in comparison to a Turbo code by itself. These results have been verified by simulation, as illustrated by the results shown in Fig. 10. An alternate technique

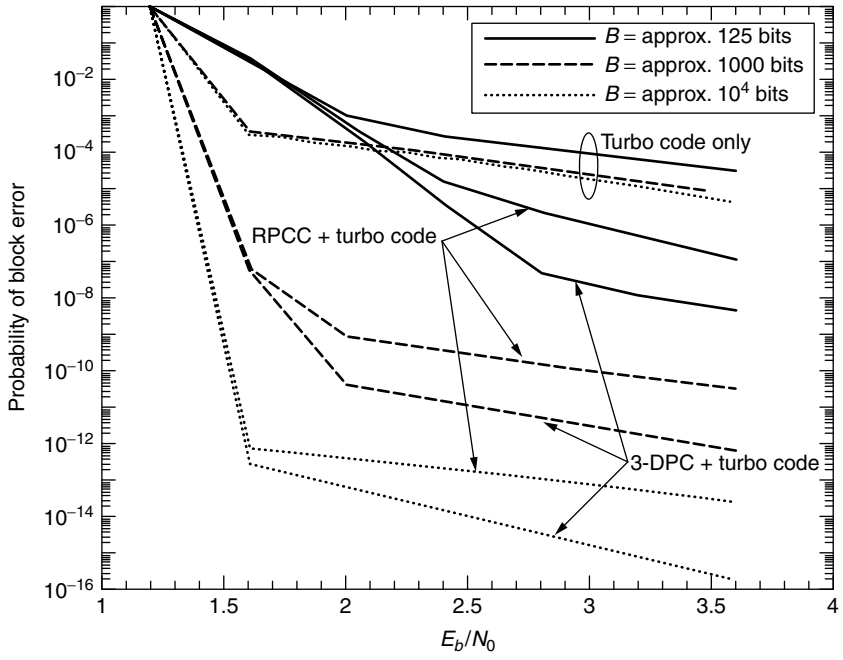


Figure 9. Bounds on the performance of concatenated outer multidimensional parity-check codes with inner rate-$\frac{1}{3}$ Turbo codes.

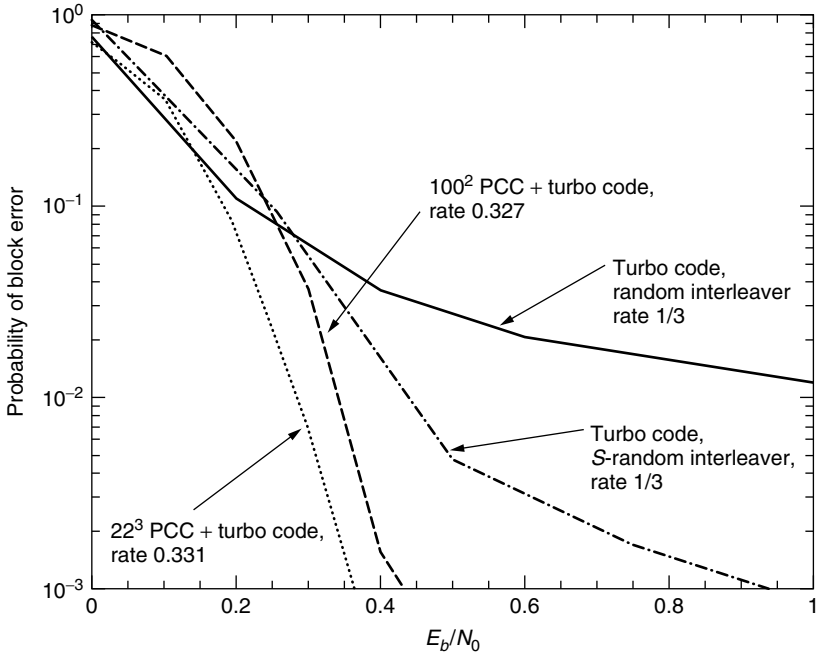


Figure 10. MDPC + Turbo codes and Turbo codes with blocklength of approximately 10^4 bits and rates approximately equal to $\frac{1}{3}$.

to improve the error floor is the use of *S*-random interleaving [40]. The simulation results presented in Fig. 10 show that the concatenated multidimensional parity-check and Turbo codes provide significantly better performance than do Turbo codes, even when *S*-random interleaving is used.

BIOGRAPHIES

John M. Shea (S'92-M'99) received the B.S. (with highest honors) in Computer Engineering from Clemson University in 1993 and the M.S. and Ph.D. degrees in Electrical Engineering from Clemson University in 1995 and 1998, respectively. In 1999 he joined the University of Florida, where he is currently an Assistant Professor of Electrical and Computer Engineering. Dr. Shea was a National Science Foundation Fellow from 1994 to 1998. He received the Ellersick Award from the IEEE Communications Society in 1996. "Dr. Shea serves as an Associate Editor for the *IEEE Transactions on Vehicular Technology*." He is currently engaged in research on wireless communications with emphasis on Turbo coding

and iterative decoding, adaptive signaling, and spread-spectrum communications.

Tan F. Wong received the B.Sc. degree (First Class Honors) from the Chinese University of Hong Kong, and the M.S.E.E., and Ph.D. degrees in Electrical Engineering from Purdue University in 1991, 1992, and 1997, respectively. He is currently an Assistant Professor of Electrical and Computer Engineering at the University of Florida. Prior to that, he was a research engineer working on the high-speed wireless networks project at the Department of Electronics at Macquarie University, Sydney, Australia. He also served as a Postdoctoral Research Associate in the School of Electrical and Computer Engineering at Purdue University. Dr. Wong serves as an Editor for the *IEEE Transactions on Communications* and as an Associate Editor for the *IEEE Transactions on Vehicular Technology*. His research interests include spread-spectrum communications, multiuser communications, error-control coding, and wireless networks.

BIBLIOGRAPHY

1. G. D. Forney, Jr., et al., Efficient modulation for bandlimited channels, *IEEE J. Select. Areas Commun.* **SAC-2**: 632–646 (Sept. 1984).
2. L.-F. Wei, Trellis-coded modulation with multidimensional constellations, *IEEE Trans. Inform. Theory* **IT-33**: 483–501 (July 1987).
3. E. Biglieri, D. Divsalar, P. J. McLane, and M. K. Simon, *Introduction to Trellis-Coded Modulation with Applications*, Macmillan, New York, 1991.
4. I. Blake, C. Heegard, T. Høholdt, and V. K.-W. Wei, Algebraic-geometry codes, *IEEE Trans. Inform. Theory* **44**: 2596–2618 (Oct. 1998).
5. K. A. S. Abdel-Ghaffar, R. J. McEliece, and H. C. A. van Tilborg, Two-dimensional burst identification codes and their use in burst correction, *IEEE Trans. Inform. Theory* **34**: 494–504 (May 1988).
6. W. J. van Gils, Two-dimensional dot codes for product identification, *IEEE Trans. Inform. Theory* **IT-33**: 620–631 (Sept. 1986).
7. P. Elias, Error-free coding, *IRE Trans. Inform. Theory* **IT-4**: 29–37 (Sept. 1954).
8. S. Lin and D. J. Costello, *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
9. H. O. Burton and E. J. Weldon, Jr., Cyclic product codes, *IEEE Trans. Inform. Theory* **IT-11**: 433–439 (July 1965).
10. W. W. Peterson and E. J. Weldon, Jr., *Error-correcting Codes*, MIT Press, Cambridge, MA, 1972.
11. L. Calabi and H. G. Haefeli, A class of binary systematic codes correcting errors at random and in bursts, *IRE Trans. Circuit* (special supplement to **CT-6**): 79–94 (May 1959).
12. E. N. Gilbert, A problem in binary encoding, *Proc. Symp. Applied Math.*, 1960, Vol. 10, pp. 291–297.
13. P. G. Neumann, A note on Gilbert burst-correcting codes, *IEEE Trans. Inform. Theory* **IT-11**: 377–384 (July 1965) (The reader should note that some of the results in this reference were corrected in Ref. 14, below).
14. L. R. Bahl and R. T. Chien, On Gilbert burst-error-correcting codes, *IEEE Trans. Inform. Theory* **IT-15**: 431–433 (May 1969).
15. R. G. Gallager, *Low-Density Parity-Check Codes*, MIT Press, Cambridge, MA, 1963.
16. L. R. Bahl and R. T. Chien, Multiple-burst-error correction by threshold decoding, *Inform. Control* **15**: 397–406 (Nov. 1969).
17. L. R. Bahl and R. T. Chien, Single- and multiple-burst-correcting properties of a class of cyclic product codes, *IEEE Trans. Inform. Theory* **IT-17**: 594–600 (Sept. 1971).
18. G. Battail, M. C. deCouvelaere, and P. Godlewski, Replication decoding, *IEEE Trans. Inform. Theory* **IT-25**: 332–345 (May 1979).
19. J. Hagenauer, E. Offer, and L. Papke, Iterative decoding of binary block and convolutional codes, *IEEE Trans. Inform. Theory* **42**: 429–445 (March 1996).
20. R. M. Pyndiah, Near-optimum decoding of product codes: Block turbo codes, *IEEE Trans. Commun.* **46**: 1003–1010 (Aug. 1998).
21. D. M. Rankin and T. A. Gulliver, Single parity check product codes, *IEEE Trans. Commun.* **49**: 1354–1362 (Aug. 2001).
22. T. F. Wong and J. M. Shea, Multi-dimensional parity check codes for bursty channels, *Proc. 2001 IEEE Int. Symp. Information Theory*, Washington, DC, June 2001, p. 123.
23. R. E. Blahut, *Principles and Practices of Information Theory*, Addison-Wesley, Reading, MA, 1987.
24. J. Chen and R. J. McEliece, *Frequency-Efficient Coding with Low-Density Generator Matrices*, Technical Report, California Institute of Technology, available at *http://www.ee.caltech.edu/systems/jfc/publications.html*.
25. D. Divsalar, S. Dolinar, F. Pollara, and R. McEliece, *Transfer Function Bounds on the Performance of Turbo Codes*, Technical Report TDA Progress Report 42-122, NASA Jet Propulsion Laboratory, Aug. 1995.
26. K. Koike and H. Ogiwara, Application of turbo codes for impulsive noise channels, *IEICE Trans. Fund.* **E81-A**: 2032–2039 (Oct. 1998).
27. E. Biglieri, J. Proakis, and S. Shamai, Fading channels: Information-theoretic and communications aspects, *IEEE Trans. Inform. Theory* **44**: 2619–2692 (Oct. 1998).
28. C. Berrou, A. Galvieux, and P. Thitimajshima, Near Shannon limit error-correcting coding and decoding, *Proc. 1993 IEEE Int. Conf. Communications*, Geneva, Switzerland, Vol. 2, 1993, pp. 1064–1070.
29. J. D. Andersen, "Turbo" coding for deep space applications, *Proc. 1995 IEEE Int. Symp. Information Theory*, Whistler, British Columbia, Canada, Sept. 1995, p. 36.
30. J. D. Andersen, Turbo codes extended with outer BCH code, *IEE Electron. Lett.* **32**: 2059–2060 (Oct. 1996).
31. K. R. Narayanan and G. L. Stüber, Selective serial concatenation of turbo codes, *IEEE Commun. Lett.* **1**: 136–139 (Sept. 1997).
32. H. C. Kim and P. J. Lee, Performance of turbo codes with a single-error correcting BCH outer code, *Proc. 2000 IEEE Int. Symp. Information Theory*, Sorrento, Italy, June 2000, p. 369.
33. O. Y. Takeshita, O. M. Collins, P. C. Massey, and D. J. Costello, Jr., On the frame-error rate of concatenated turbo codes, *IEEE Trans. Commun.* **49**: 602–608 (April 2001).

34. D. J. Costello, Jr. and G. Meyerhans, Concatenated turbo codes, *Proc. 1996 IEEE Int. Symp. Information Theory and Applications*, Victoria, Canada, Sept. 1996, pp. 571–574.

35. M. C. Valenti, Inserting turbo code technology into the DVB satellite broadcast system, *Proc. 2000 IEEE Military Communications Conf.*, Los Angeles, Oct. 2000, pp. 650–654.

36. J. M. Shea, Improving the performance of turbo codes through concatenation with rectangular parity check codes, *Proc. 2001 IEEE Int. Symp. Information Theory*, Washington, DC, June 2001, p. 144.

37. J. M. Shea and T. F. Wong, Concatenated codes based on multidimensional parity-check codes and turbo codes, *Proc. 2001 IEEE Military Communications Conf.*, Washington, DC, Oct. 2001, Vol. 2, pp. 1152–1156.

38. 3rd Generation Partnership Project, *Technical Specification TS 25.212 v4.1.0: Radio Access Network: Multiplexing and Channel Coding* (*FDD*), Technical Report, available on the Web at *ftp://ftp.3gpp.org/Specs/2001-06/Rel-4/25_series/25212-410.zip* (June 2001).

39. 3rd Generation Partnership Project 2, *Physical Layer Standard for cdma2000 Spread Spectrum Systems-Release 0-version 3.0*, Technical Report, available on the Web at *http://www.3gpp2.org/Public_html/specs/C.S0002-0_v3.0.pdf* (July 2001).

40. S. Dolinar and D. Divsalar, *Weight Distributions for Turbo Codes Using Random and Nonrandom Permutations*, Technical Report TDA Progress Report 42-122, NASA Jet Propulsion Laboratory, Aug. 1995.

MULTIMEDIA MEDIUM ACCESS CONTROL PROTOCOLS FOR WDM OPTICAL NETWORKS

MOUNIR HAMDI
Hong Kong University of Science and Technology
Hong Kong

MAODE MA
Nanyang Technological University
Singapore

1. INTRODUCTION

Wavelength-division multiplexing (WDM) is the most promising multiplexing technology for optical networks. By using WDM, the optical transmission spectrum is configured into a number of nonoverlapping wavelength bands. In particular, multiple WDM channels could be allowed to coexist on a single fiber. As a result, the requirements on balancing the optoelectronic bandwidth mismatch could be met by designing and developing appropriate WDM optical network architectures and protocols.

WDM optical networks can be designed using one of two types of architecture: broadcast-and-select networks or wavelength-routed networks. Typically, the former is used for local-area networks, while the latter is used for wide-area networks. A local WDM optical network may be set up by connecting computing nodes via two-way fibers to a passive star coupler, as shown in Fig. 1. A node can send its information to the passive star coupler on one available wavelength by using a laser, which produces an optical stream modulated with information. Modulated optical streams from transmitting nodes are combined by the passive star coupler. Then the integrated stream is separated and transmitted to all the nodes in the network. A destination's receiver is an optical filter and is tuned to one of the wavelengths to receive its designated information stream. Communication between the source and destination nodes is implemented in one of the following two modes: *single-hop*, in which communication takes place directly between two nodes [1], or *multihop*, in which information from a source to a destination may be routed through the intermediate nodes of the network [2].

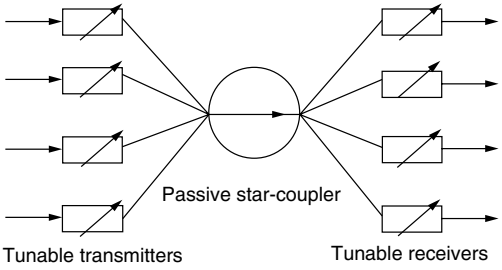


Figure 1. A single-hop passive star coupled WDM optical network.

On the basis of the architectures of WDM optical networks, medium access control (MAC) protocols are required to efficiently allocate and coordinate the system resources. The challenge of developing appropriate MAC protocols is to efficiently exploit the potential vast bandwidth of an optical fiber to meet the increasing information transmission demand under the constraints of the network resources and the constraints imposed on the transmitted information.

In this article, we will review state-of-the-art MAC protocols in passive star coupler–based WDM networks. Because the single-hop structure has more dynamics in nature than the multihop one, we will focus on the MAC protocols for the single-hop architecture. We will also discuss several protocols in some detail to show their importance in the development of MAC protocols for the single-hop passive star coupler–based WDM networks. According to the network service provided to the transmitted information, we roughly divide the MAC protocols into three categories as follows: MAC protocols for packet and variable-length message transmission, MAC protocols for real-time message transmission (including MAC protocol with QoS concerns), and MAC protocols for multimedia applications.

The remainder of this article is organized as follows. Section 2 reviews the MAC protocols of transmission service for packets and variable-length messages. Section 3 discusses the MAC protocols for real-time service. Section 4 provides an investigation on multimedia protocols. Section 5 concludes the article with a summary.

2. MAC PROTOCOLS FOR PACKET AND VARIABLE-LENGTH MESSAGE TRANSMISSION

Numerous MAC protocols for normal data transmission have been proposed since 1990 for the WDM single-hop network architecture. According to how the data are

presented and transmitted in the networks, the MAC protocols can be simply grouped as MAC protocols for fixed-size packet transmission and MAC protocols for variable-size message transmission.

2.1. MAC Protocols for Packet Transmission

The MAC protocols for packet transmission in single-hop passive star–coupled WDM networks are so called "legacy" protocols because they are dedicated for fixed-length packet transmission, and they are often adopted from legacy shared medium networks. In a single-hop network, a significant amount of dynamic coordination between nodes is required in order to access the network resources. According to the coordination schemes, the MAC protocols can be further classified into the following subcategories.

2.1.1. Nonpretransmission Coordination Protocols. Protocols with nonpretransmission coordination do not have to reserve any channels for pretransmission coordination. All the transmission channels are either preassigned to transmitting nodes or accessed by transmitting nodes through contest. These protocols can be categorized accordingly in the following subgroups.

2.1.1.1. Fixed Assignment. A simple approach, based on the fixed-wavelength assignment technique, is time-division multiplexing (TDM) extended over a multichannel environment [3]. It is predetermined that a pair of nodes is allowed to communicate with each other in the specified time slots within a cycle on the specified channel. Several extensions to the abovementioned protocol have been proposed to improve the performance. One approach, *weighted TDM*, assigns a different number of time slots to different transmitting nodes according to the traffic load on each node [4]. Another proposed approach is a versatile time–wavelength assignment algorithm [5]. Under the condition that a traffic demand matrix is given beforehand, the algorithm can minimize the tuning times and has the ability to reduce transmission delay. Some new algorithms based on the abovementioned algorithm [5] study problems such as the performance of scheduling packet transmissions with an arbitrary traffic matrix and the effect of the tuning time on the performance [6–8].

2.1.1.2. Partial Fixed Assignment Protocols. Three partial fixed assignment protocols have been proposed [3]. The first one is the destination allocation (DA) protocol. By using this protocol, the number of source and destination node pairs can be the same as the number of nodes. A source allocation (SA) protocol is also defined in which the control of access to transmission channels is further relaxed. Similar to the SA protocol, an allocation-free (AF) protocol has been proposed, in which all source–destination pairs of computing nodes have full rights to transmit packets on any channel over any time slot duration.

2.1.1.3. Random Access Protocols. Two slotted-ALOHA protocols have been proposed [9]. Using the first protocol, time is slotted on all transmission channels, and these slots are synchronized across all channels. Using the second protocol, each packet can have several numbers of minislots, and time across all channels is synchronized over minislots. In addition, two similar protocols appeared in the literature [10].

2.1.2. Pretransmission Coordination Protocols. Employing protocols that do require pretransmission coordination, transmission channels are grouped into control channels and data channels. These protocols can be categorized according to the ways to access the control channels into the following subgroups:

2.1.2.1. Random-Access Protocols. The architecture of the network protocols in this subgroup is as follows. In a single-hop communication network, a control channel is employed. Each node is equipped with a single tunable transmitter and a single tunable receiver.

Habbab et al. [11] describe three random-access protocols such as ALOHA, slotted ALOHA, and CSMA are proposed to access the control channel. ALOHA, CSMA, and the N-server switch scheme can be the subprotocols for the data channels. Under a typical ALOHA protocol, a node transmits a control packet over the control channel at a randomly selected time, after which it immediately transmits a data packet on a data channel, which is specified by the control packet.

Mehravari [12] has proposed an improved protocol, slotted-ALOHA/delayed-ALOHA. This protocol requires that a transmitting node delay transmitting data on a data channel until it gets the acknowledgment that its control packet has been successfully received by the destination node. The probability of data channel collisions can be decreased, and the performance in terms of throughput can be improved.

Sudhakar et al. [13] proposed one set of slotted-ALOHA protocols and one set of reservation-ALOHA protocols. The set of the slotted-ALOHA-based protocols are improvements over the protocols proposed by Habbab et al. [11].

A so-called multicontrol channel protocol has been proposed [14] that aims at improving reservation-ALOHA-based protocols. All channels are used to transmit control information as well as data information. Control packet transmission uses a contention-based operation; while data transmission follows it.

These protocols basically cannot prevent receiver collisions. A protocol that is especially designed to avoid receiver collision has been proposed [15].

2.1.2.2. Reservation Protocols. Using the dynamic time–wavelength-division multiple-access (DT-WDMA) protocol, a channel is reserved as a control channel and it is accessed only in a preassigned TDM fashion. It requires that each node have two transmitters and two receivers [16]. One pair of the transceivers is fixed to the control channel, while another pair is tunable to all the data channels. If there are N nodes in the network, N data channels and one control channel are required. Although this protocol cannot avoid receiver collisions, it ensures that exactly one data item can be successfully accepted when more than one data packet come to the same destination node simultaneously.

One proposal [17] to improve the TD-WDMA algorithm is to use an optical delay line to buffer the potential collided packets, when more than one node transmits data packets to the same destination node at the same time. Its effectiveness depends on the relative capacity of the buffer. Another protocol [18] also tries to improve the TD-WDMA algorithm by making transmitting nodes remember the information from the previous transmission of a control packet and combining this information into the scheduling of packet transmission.

Another two protocols [19,20] intended to improve the TD-WDMA algorithm are outlined. The first one is called the *dynamic allocation scheme* (DAS), where each node runs an identical algorithm based on a common random number generator with the same seed. The second protocol is termed *hybrid TDM*. Time on the data channels is divided into frames consisting of several slots. In a certain period of time, one slot will be opened for a transmitting node to transmit data packets to any destination receiver.

A reservation-based multicontrol channel protocol has been described [21]. Employing this protocol, x channels $[1 < x < (N/2)]$ can be reserved as control channels to transmit control information, where N is the number of channels in the network. The value of x is a system design parameter, which depends on the ratio of the amount of control information and the amount of actual data information. The objective to reserve multiple control channels in the network is to decrease the overhead of control information processing time as much as possible.

The properties of the "legacy" MAC protocols can be summarized based on the basis of the abovementioned survey as follows. Although the protocols using the fixed-channel assignment approach can ensure that data are successfully transmitted and received, they are sensitive to the dynamic bandwidth requirements of the network and are difficult to scale in terms of the number of nodes. The protocols using the contention-based channel assignment approach introduce contention on data channels in order to adapt to the dynamic bandwidth requirements. As a result, either channel collision or receiver collision will occur. The protocols with contention-based control channel assignment still have either a data channel collision or a receiver collision because contention is involved in the control channel. Some protocols [22,23] have the capability to avoid both collisions by continuously testing the network states. The reservation-based protocols, which use a fixed control channel assignment approach, can only ensure data transmission without collisions. However, by introducing some information to make the network nodes intelligent, it has the potential to avoid receiver collisions as well. It also has the potential to accommodate application traffic composed of variable-length messages.

2.2. MAC Protocols for Variable-Length Message Transmission

The "legacy" MAC protocols are designed to handle and schedule fixed-length packets. Using these MAC protocols, most of the application-level data units (ADUs) must be segmented into a sequence of fixed-size packets for transmission over the networks. However, as traffic streams in the real world are often characterized as bursty, consecutive arriving packets in a burst are strongly correlated by having the same destination node. An intuitive idea about this observation is that all the fixed-size packets of a burst should be scheduled as a whole and transmitted continuously in a WDM network rather than be scheduled on a packet-by-packet basis. Another way of looking at this is that the ADUs should not be segmented. Rather, they should be simply scheduled as a whole without interleaving. The main advantages of using a burst-based or message transmission over WDM networks are (1) to an application, the performance metrics of its data units are more relevant performance measures than ones specified by individual packets; (2) it perfectly fits the current trend of carrying IP traffic over WDM networks; and (3) message fragmentation and reassembly are not needed.

The first two MAC protocols proposed by Sudhakar et al. [13] for variable-length message transmission are protocols with contention-based control channel assignment. Another two reservation-ALOHA-based protocols [13] are presented in order to serve the long-holding-time traffic of variable-length messages. The first protocol aims to improve the basic slotted-ALOHA-based technique. The second protocol aims to improve the slotted-ALOHA-based protocol with asynchronous cycles on the different data channels. Data channel collisions can be avoided by the two reservation-ALOHA-based protocols.

Another protocol [24–26] tries to improve the reservation-based TD-WDMA protocol [16]. The number of nodes is larger than the number of channels; the transmitted data are in the form of a variable-length message rather than a fixed-length packet; data transmission can start without any delay. Both data collision and receiver collision can be avoided because any message transmission scheduling has to consider the status of the data channels as well as receivers.

Two other protocols, FatMAC [27] and LiteMAC [28], try to combine reservation-based and preallocation-based techniques to schedule variable-length message transmission. FatMAC is a hybrid approach that reserves access to preallocated channels through control packets. Transmission is organized into cycles where each of them consists of a reservation phase and a data phase. A reservation specifies the destination, the channel and the message length of the next data transmission. The LiteMAC protocol is an extension of FatMAC. Using the LiteMAC protocol, each node is equipped with a tunable transmitter and a tunable receiver rather than a fixed receiver as in FatMAC. LiteMAC has more flexibility than FatMAC because of the usage of a tunable receiver and its special scheduling mechanism. Hence, more complicated scheduling algorithms could be used to achieve better performance than FatMAC. Both FatMAC and LiteMAC have the ability to transmit variable-length messages by efficient scheduling without collisions. The performances of these two protocols have been proved to be better than that of the preallocation-based protocols, while fewer transmission channels are used than in reservation-based protocols. With these two protocols, low average message delay and high channel utilization can be expected.

2.2.1. A Reservation-Based MAC Protocol for Variable-Length Messages. Proposed [29], based on the protocol advanced by Bogineni and Dowd Jia et al. [24–26]; this was an intelligent reservation-based protocol for scheduling variable-length message transmission. The protocol employs some global information of the network to avoid both data channel collisions and receiver collisions while message transmission is scheduled. Its ability to avoid both collisions makes this protocol a milestone in the development of MAC protocols for WDM optical networks.

The network consists of M nodes and $W+1$ WDM channels. W channels are used as data channels. The other channel is the control channel. Each node is equipped with a fixed transmitter and a fixed receiver for the control channel, and a tunable transmitter and a tunable receiver to access the data channels. The time on the data channels is divided into data slots. It is assumed that there is a networkwide synchronization of data slots over all data channels. The duration of a data slot is equal to the transmission time of a fixed-length data packet. A node generates variable-length messages, each of which contains one or more fixed-length data packets. On the control channel, time is divided into control frames. A control frame consists of M control slots. A control slot has several fields such as address of destination node and the length of the message. A time-division multiple-access protocol is employed to access the control channel so that the collision of control packets can be avoided.

Before a node sends a message, it needs to transmit a control packet on the control channel in its control slot. After one round-trip propagation delay, all the nodes in the network will receive the control packet. Then a distributed scheduling algorithm is invoked at each node to determine the data channel and the time duration over which the message will be transmitted. Once a message is scheduled, the transmitter will tune to the selected data channel and transmit the scheduled message at the scheduled transmission time. When the message arrives at its destination node, the receiver should have been tuned to the same data channel to receive the message.

The data channel assignment algorithm determines the data channel and the time duration over which the message will be transmitted. The algorithm schedules message transmissions based on some global information in order to avoid the data channel collisions and the receiver collisions. The global information is expressed through two tables, which reside at each node. One table is the *receiver available-time table* (RAT). RAT is an array of M elements, one for each node. $\mathrm{RAT}[i]=n$, where $i=1,2,\ldots,M$, means that node i's receiver will become free after n data slots. If $n=0$, then node i's receiver is currently idle, and no reception is scheduled for it as yet. RAT is needed for avoiding receiver collisions. Another table is the *channel available-time table* (CAT). CAT is an array of W elements, one for each data channel. $\mathrm{CAT}[k]=m$, where $k=1,2,\ldots,W$, means that data channel k will be available after m data slots. If $\mathrm{CAT}[k]=0$, data channel k is currently available. CAT is needed to avoid collisions on data channels. Local and identical copies of these two tables are at each node. They contain consistent information on the messages whose transmissions have been scheduled but not yet transmitted. The contents of the tables are relative to current time. Three data channel assignment algorithms have been proposed. The fundamental one, the *earliest available-time scheduling* (EATS) algorithm. This algorithm schedules the transmission of a message by selecting a data channel, which is the earliest available.

This reservation-based protocol has been shown to have quite good performance while it can avoid data channel collisions and receiver collisions.

2.2.2. A Receiver-Oriented MAC Protocol for Variable-Length Messages. Some related protocols have been proposed to improve the performance of the network based on the same system architecture of Ref. 29. In Ref. 30, the proposed protocol tries to avoid the head-of-queue blocking during the channel assignment procedure by introducing the concept of "destination queue" to make each node maintain M queues, where M is the number of nodes in the network. Hamidzadeh et al. [31], notice that the performance of the network could be further improved by the way of exploiting more existing global information of the network and the transmitted messages. From this point of view, a general scheduling scheme, which combines the message sequencing techniques with channel assignment algorithms [29], is proposed to schedule variable-length message transmission. In Ref. 32, as an example of the general scheduling scheme in Ref. 31, a new scheduling algorithm is proposed. This algorithm, *receiver-oriented earliest available-time scheduling* (RO-EATS), decides the sequence of the message transmission using the information of the receiver's states to decrease message transmission blocking caused by avoiding collisions.

The RO-EATS scheduling algorithm employs the same system structure and network service as those of the protocol in [29] to form a receiver-oriented MAC protocol, which is an extension of the protocol [29]. The logic structure of the system model for the RO-EATS protocol can be expressed as in Fig. 2. Employing the new protocol, the management of messages' transmission and reception is the same as that of the Jia et al. protocol [29]. The difference between the two protocols is in the scheduling algorithm for message transmission. The RO-EATS algorithm works as follows. It first considers the

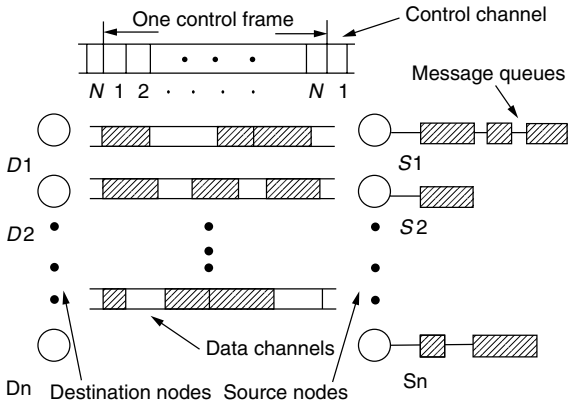


Figure 2. The network architecture for variable-length messages.

earliest available receiver among all the nodes in the network and then selects a message, which is destined to this receiver from those, which are ready and identified by the control frame. After that, a channel is selected and assigned to the selected message by the principle of the EATS algorithm. The scheme to choose a suitable message to transmit is based on the information of the states of the receivers presented in RAT. The objective of this scheme is to avoid lots of messages going to one or a few nodes at the same time and try to raise the channels utilization. The motivation of the algorithm comes from the observation that two consecutive messages with the same destination may not fully use the available channels when the EATS algorithm is employed. The new algorithm enforces the idea of scheduling two consecutive messages away from going to the same destination node. The RO-EATS algorithm always checks the table of RAT to see which node is the least visited destination and to choose the message, which is destined to this node to transmit. In this way, the average message delay can be shown to be quite low and channel utilization can be shown to be high.

3. MAC PROTOCOLS FOR REAL-TIME SERVICE

An important function of high-speed computer networks such as WDM optical networks is to provide real-time service to time-constrained application streams such as video or audio information. Most of the MAC protocols that provide real-time service on passive star–coupled WDM optical networks are protocols with reservation-based precoordination. According to the type of the real-time service provided to the transmitted messages, the MAC protocols for real-time service can be simply classified into two types: protocols with best-effort service and protocols with quality-of-service (QoS) capabilities.

3.1. MAC Protocols for Best-Effort Real-Time Service

A protocol termed *time-deterministic time- and wavelength-division multiple access* (TD-TWDMA) [33] provides services for both hard real-time messages and soft real-time messages for single destination, multicast, and broadcast transmissions. All channels can be accessed by a fixed-assignment method, which is a TDM approach. Using this approach, each channel is divided into time slots. Each node has a number of slots for hard real-time message transmission. Soft real-time messages can be transmitted if there is no hard real-time message requiring service. Each node is equipped with one fixed transmitter and tunable receivers. The transmitter is fixed to its assigned channel, while the receiver can be tuned over all channels in the network. Each node has a specified channel because the number of nodes, C, is equal to the number of channels, M, in the network. At each node, there are $2 * M$ queues, M queues for the hard real-time messages, and another M queues for the soft real-time messages. For each type of queue, one queue is for broadcast and $M - 1$ queues for the single destination. The messages in the broadcast queue can be either control information or data to be broadcast. The protocol works as follows. First it sends a broadcast slot containing the control information; then it invokes the slot allocation algorithm to determine the slots used to transmit the data information; Finally, each node tunes to the specified channel to receive the data. The slot allocation algorithm follows the static priority approach. The basic idea of the algorithm can be summarized as follows: (1) the M hard real-time message queues have higher priorities; while the M soft real-time message queues have lower priorities; (2) each queue in each group has a fixed priority, while the queues for broadcasting have the highest priority in each group; (3) message transmission scheduling is based on the queue priority; and (4) for the hard real-time messages, if transmission delay is over their deadlines, these messages will be dropped; while for the soft real-time messages, they will be scheduled whether they are beyond their deadlines or not.

A reservation-based MAC protocol for best-effort real-time service can be found in the literature [34]. This protocol is for the same network architecture as that in Jia et al. [29]. Both hard real-time and soft real-time variable-length message transmissions have been considered. The scheduling algorithms of the protocol are based on the time-related dynamic priority scheme, namely, *minimum laxity first* (MLF) scheduling. The principle of this dynamic scheduling scheme is that the most stringent message will get the transmission service first. This protocol employs global information of the network as well as the transmitted messages to ensure zero message loss rate caused by both data channel collisions and receiver collisions and decrease the message loss rate caused by network delay. This research work has confirmed that when real-time traffic is introduced in the networks, dynamic time-based priority assignment schemes as well as priority-based scheduling algorithms should be employed to improve the real-time performance of the networks as much as possible.

A novel reservation-based MAC protocol for real-time service has been proposed [35] that extends the functions of the protocol in Ma et al. [34] to provide differentiated service to benefit both real-time and non-real-time applications in one topology. This protocol considers the transmission of the variable-length messages with hard real-time constraints, soft real-time constraints, or non-real-time constraints. The scheduling algorithm, *minimum laxity first with time tolerance scheduling* (MLF-TTS), *algorithm*, of this protocol schedules real-time message transmission according to their time constraints. The basic minimum laxity first (MLF) scheduling policy is adopted for scheduling real-time traffic. For non real-time messages, the scheduling algorithm manages their transmission based on the following fact of the transmission of real-time messages. After the real-time messages have been scheduled for transmission on certain channels in certain time slots to their destination nodes, some of them could be blocked just because there may be more than two consecutive messages going to the same destination node in a very short time period. This fact causes the utilization of the transmission channels to be quite low, and succeeding messages will be blocked so that the average message delay for the non-real-time messages will be very high. The MLF-TTS algorithm seeks and takes a time period, in which the real-time messages are

being blocked to wait for their destinations to be free, to schedule the transmission of non-real-time messages under the condition that the transmission time of these messages should be less than the time that the blocked real-time messages are waiting for their destinations to be available. Since the global information of the receivers and channels in the network are available to every source node, this scheduling is feasible and can be easily implemented. Using the MLF-TTS algorithm, the average message delay for the messages without time constraints could be expected to decrease while the message loss rate or message tardy rate is kept as low as those of the simple MLF algorithms. In addition, the channel utilization could be expected to be high. Unlike the scheduling algorithms, which aim to only decrease the average message delay, the MLF-TTS could be expected to significantly increase the real-time performance of the WDM MAC protocols. As a result, a fairness transmission service to both real-time and non real-time traffics could be achieved.

3.2. MAC Protocols with Quality-of-Service Concerns

Quality of service (QoS) is an important issue when real-time applications demand a given network transmission service. It is obvious that the QoS provided by a network service to real-time applications indicates the degree to which the real-time applications can meet their time constraints. However, best-effort real-time network service cannot ensure QoS because it cannot guarantee that real-time applications can meet their time constraints to a certain degree when they are transmitted. It is necessary to develop MAC protocols with QoS capabilities so that the QoS guaranteed by the network service could be estimated and predicted. There are two types of MAC protocols with QoS capabilities: protocols with deterministically guaranteed service and protocols with statistically guaranteed service.

3.2.1. MAC Protocols with Deterministically Guaranteed Service. A preallocation-based channel access protocol has been proposed [36] to provide deterministic timing guarantees to support time-constrained communication in a single-hop passive star−coupled WDM optical network. This protocol takes a passive star−coupled broadcast-and-select network architecture in which N stations are connected to a passive star coupler with W different wavelength channels. Each W channel is slotted and shared by the N stations by means of a TDM approach. The slots on each channel are preassigned to the transmitters. A schedule specifies, for each channel, which slots are used for data transmission from node i to node j, where $1 \le i \le N,\ 1 \le j \le N,\ i \le j$. Each node of the network can be equipped with a pair of tunable transmitters and tunable receivers, which can be tuned over all the wavelengths. Each real-time message stream with source and destination nodes specified is characterized with two parameters, relative message deadline D_i and maximum message size C_i, which can arrive within any time interval of length D_i. A scheme called a *binary splitting scheme* (BSS) is proposed to assign each message stream sufficient and well-spaced slots to fulfill its timing requirement. Given a set of real-time message streams M specified by the maximum length of each stream C_i and the relative deadline of each stream D_i, this scheme can allocate time slots over as few channels as much as possible in such a way that at least C_i slots are assigned to M_i in any time window of size D_i slots so that the real-time constraints of the message streams can be guaranteed.

A modified preallocation-based MAC protocol has been proposed [37] to guarantee a reserved bandwidth and a constant delay bound to the integrated traffic. This protocol works as a centralized scheduler based on the star-coupled broadcast LAN topology, which is similar to that proposed by Jia et al. [29]. Each node in the LAN is equipped with a pair of tunable transceivers. The access to the transmission channels is controlled by the scheduler, which is based on the concept of computing maximal weighted matching, a generalization of maximal matching on an unweighted graph. According to this concept, several scheduling algorithms have been proposed. A *credit-weighted algorithm* is proposed to serve guaranteed traffic. A *bucket-credit-weighted algorithm* is designed to serve bursty traffic, and a *validated queue algorithm* is a modification of the *bucket-credit-weighted algorithm* to serve bursty traffic and keep throughput guarantee at the same time. It has been proved that these scheduling algorithms can guarantee the bandwidth reservation to a certain percentage of the network capacity and ensure a small delay bound even when bursty traffic exists.

A reservation-based MAC protocol for deterministic guaranteed real-time service has been proposed [38]. This protocol is for the same network structure as that in Jia et al. [29]. In this protocol [38], a systematic scheme is proposed to provide deterministic guaranteed real-time service for application streams composed of variable-length messages. It includes an admission control policy, traffic regularity, and message transmission scheduling algorithm. A traffic-intensity-oriented admission control policy is developed to manage flow-level traffic. A g-regularity scheme based on the max-plus algebra theory is employed to shape the traffic. An *adaptive round-robin and earliest available time scheduling* (ARR-EATS) algorithm is proposed to schedule variable-length message transmission. All of these are integrated to ensure that a deterministic guaranteed real-time service can be achieved.

3.2.2. MAC Protocols for Statistically Guaranteed Service. The MAC protocols with deterministically guaranteed service can normally guarantee specific transmission delays to real-time applications, or, under certain time constraints imposed to the real-time applications, a specific percentage of real-time messages, which can meet the time constraints, can be predicted. However, MAC protocols for statistically guaranteed service cannot provide a deterministic guaranteed QoS service. Only an estimated percentage of real-time messages, which can meet their time constraints, can be evaluated statistically. Most of the MAC protocols in this category consider the issue of providing differentiated service to both real-time and non-real-time applications. Using these protocols, statistical QoS to real-time applications can be expected by sacrificing the transmission service to non-real-time applications.

A reservation-based protocol has been proposed [39] to provide statistically guaranteed real-time services in WDM optical token LANs. In the network, there are M nodes and $W+1$ channels. One of the channels is the control channel, while the others are data channels. Different from the network structure presented by Jia et al. [29], the control channel in this network is accessed by token passing. At each node, there is a fixed receiver and transmitter tuned to the control channel. There is also a tunable transmitter, which can be tuned to any of the data channels. There are one or more receivers fixed to certain data channels. The protocol provides transmission service to either real-time or non-real-time messages. The packets in the traffic may have variable length but be bounded by a maximum value. At each node, there are W queues, each of which corresponds to one of the channels in the network. The messages come into one of the queues according to the information of their destination nodes and the information of the channels, which connect to the corresponding destination nodes. The protocol works as follows. A token exists on the control channel to ensure collision-free transmission on data channels. The token has a designated node K. Every node can read the contents of the token and updates its local status table by the information in the fields of the token. When node K observes the token on the network, it will check the available channels. If there are no channels available, node K gives up this opportunity to send its queued packets. Otherwise, the *priority index algorithm* (PIA) is invoked to evaluate the priority of each message queue on node K and then uses the *transmitter scheduling algorithm* (TSA) to determine the transmission channel. Also the *flying-target algorithm* (FTA) is used to decide the next destination of the control token. After all these have completed, node K's status and scheduling result will be written into the token. Then the token on node K will be sent out, and the scheduled packets will be transmitted.

A novel reservation-based MAC protocol has been proposed [40] to support statistically guaranteed real-time service in WDM networks by using a hierarchical scheduling framework. This work is developed for a network structure similar to that described by Jia et al. [29]. The major advantage of its protocol over that proposed by Yan et al. [39] is that it divides the scheduling issue into flow scheduling or VC scheduling and transmission scheduling. The former is responsible for considering the order of traffic streams to be transmitted. It schedules packets to be transferred from VC queues to the transmission queue. The latter is to decide the order of the packets transmission. The packets involved in the transmission scheduling are those selected from the traffic streams by the flow schedule scheme. A simple-round robin scheme is adopted in the VC scheduling, and a random scheduling with age priority is used in the transmission scheduling. Another good point of this protocol is that a rescheduling scheme is employed to compensate the failure scheduling result due to either output conflict or channel conflict. Using this scheme, if a scheduling fails, a decision has to be made as to whether rescheduling the same packet or scheduling a new packet from another VC is performed. If the failure is from a real-time traffic, it certainly makes sense to reschedule the very same packet as soon as possible. The very same real-time packet will be retransmitted immediately in the next control slot; thus no other new scheduling either from real-time traffic or non-real-time traffic of the same source node can be initiated. The more intriguing part of the scheduling algorithm is the rescheduling of non-real-time traffic. If real-time traffic has more stringent QoS requirements, the rescheduling scheme will ignore rescheduling the failed non-real-time packet to ensure that the real-time traffic meets its time constraints. Compared with the protocol proposed by Yan et al. [39], this protocol is expected to diminish the ratio of the packet, which are over their deadlines.

A protocol similar to that [39] has been proposed [41]. This protocol is based on the same network architecture as that described by Yan et al. [39], which is a WDM optical token LAN. The protocol tries to provide fairness transmission service to both real-time and non-real-time messages in the same network, while the QoS of the real-time messages could be adjusted to a reasonable level. The protocol separates the real-time and non-real-time messages into different queues at each transmission node. The real-time message queue has higher priority for transmission than does that for non-real-time messages. The outstanding point of this protocol is that the scheduling scheme of the protocol has set up a threshold on the queue length for the non-real-time message queue in order to balance the transmission service. The operation of the scheduling scheme works as follows. When the length of the non-real-time message queue has not reached the threshold, the real-time messages will be scheduled for transmission. However, when the threshold has been reached or exceeded, the non-real-time messages will be scheduled for transmission until the length of the lower-priority queue is under the threshold. The QoS of the real-time message transmission is measured by the loss rate. With setting of the threshold, the scheduling scheme can provide fair transmission service to both real-time and non-real-time traffic with certain QoS guarantee to real-time traffic. Alternatively, with a change of the threshold, the level of QoS guarantee to real-time traffic can be controlled.

A MAC protocol to provide statistical QoS service has been presented [42], that is based on a multichannel ring topology. This topology is somewhat different from that described by Yan et al. [39]. In this network, every node is equipped with a fixed receiver and a tunable transmitter. Every transmission channel is associated with each destination node. The transmitted information is in the form of fixed-size packets. A collision-free MAC protocol is proposed, known as *synchronous round-robin with reservation* (SR^3), to support both QoS guarantee to real-time traffic and best-effort service to non-real-time traffic. There are three components in the SR^3 protocol. The access strategy selects proper packets to transmit, the fairness control algorithm guarantees the throughput fairness among all the channels in the multiring network, and the reservation scheme allows the transmitting nodes to dynamically allocate a portion of available bandwidth. This protocol has been proved to have the capability to

provide quality of service to both real-time and non-real-time traffic in the multiring topology.

4. MAC PROTOCOLS FOR MULTIMEDIA APPLICATIONS

There has been a rapid growth in the number of multimedia applications. Different multimedia applications require various classes of transmission service, including the transmission of data, audio, and various types of video and images on WDM optical networks. High-speed protocols including the protocols at the medium access control layer are needed to cater for the different requirements of the transmission of various multimedia applications.

Multimedia applications contain a variety of media: data, graphics, images, audio, and video. The transmission of the multimedia applications is a kind of real-time and stream-oriented communication. The quality of service required of a stream communication includes guaranteed bandwidth (throughput), delay, and delay variation (jitter). However, the quality of service for different kinds of applications varies. On one hand, hard real-time traffic such as voice and video require stringent time delay and delay variance, but tolerates a small percentage of packet loss. On the other hand, soft or non-real-time traffic such as images, graphics, text, and data requires no packet loss, but tolerates time delay. Hence, the protocols that provide transmission service to multimedia applications should support and ensure the variety of QoS requirements of different types of media.

Support of multimedia applications by MAC protocols has become a hot topic in the field of research on WDM optical networks. Some research results have been generated from existing protocols for real-time service. However, some protocols are completely novel or based on new network architectures dedicated to multimedia traffic.

4.1. Modified Protocols for Multimedia Applications

The feasibility of several existing protocols based on WDM bus LAN architecture to support multimedia applications has been studied [43]. Using a simulation study, the authors point out that several currently existing MAC protocols such as FairNet, WDMA, and nDQDB are not satisfactory for supporting multimedia traffic in the sense that these protocols cannot guarantee that the total delay or jitter will not grow beyond the accepted value for different classes of multimedia applications. A further study on several MAC protocols to support multimedia traffic on WDM optical networks has been carried out [44]. These protocols, including distributed queue dual bus (DQDB), cyclic-reservation multiple access (CRMA), distributed-queue multipleaccess (DQMA), and fair distributed queue (FDQ) are distributed reservation access schemes for WDM optical networks, based on slotted unidirectional bus structures. The performance of these four protocols is studied to simultaneously support synchronous traffic (for various real-time multimedia applications) and asynchronous traffic (for interactive terminal activities and data transfers). The authors have pointed out, through extensive simulation results, that the reservation-based protocols are suitable for integrating real-time multimedia traffic with bursty data traffic in the WDM optical network when the delay constraint is somewhat relaxed. The FDQ protocol stands out for supporting heterogeneous traffic. The results from the two studies cited above imply that the reservation-based protocols have the potential to accommodate the multimedia transmission in the WDM optical network rather than the nonprecoordination-based protocols.

A video-on-demand system over WDM optical networks has been studied [45]. The video-on-demand (VOD) application is considered different from live video application or MPEG compressed video/audio application in that the live video traffic is a variable-bit-rate (VBR) traffic because the video/audio sources of the application should be captured, compressed and then transmitted in real-time fashion; while VOD traffic is a constant-bit-rate (CBR) traffic because the video/audio sources of the application are processed in advance, kept on the video server, and transmitted at a regular rate. The VOD traffic is desirable to be served by isochronous transmission service by the network. The network structure of the VOD system [45] is a passive star coupler–based WDM optical network. Each node in the network is equipped with multiple tunable transmitters and receivers for data transmission and one pair of fixed transceivers for the control channel access. A centralized medium access control scheduler is employed to schedule the isochronous and the asynchronous traffic demands. A scheduling algorithm, KT-MTR, is employed for scheduling the asynchronous traffic only. Another scheduling algorithm, IATSA-MTR, is presented for scheduling both isochronous and asynchronous traffic that coexist in the network. These scheduling algorithms are shown to be efficient for serving VOD applications in star coupler–based WDM optical networks.

In order to efficiently support various types of traffic streams with different characteristics and QoS requirements in a single WDM optical network and to dynamically allocate the network bandwidth to the different classes of traffic so that the network performance could be boosted, a novel approach to integrate different types of existing medium access control protocols into a single MAC protocol in the specified WDM network architecture has been proposed [46]. The network architecture for this unique MAC protocol is similar to that proposed by Jia et al. [29], which is a passive star coupler–based WDM optical LAN. The main difference between these architectures is that each transmitting node in the architecture described by Wang and Hamdi [46] has been equipped with three pairs of tunable transceivers for three types of traffic streams working in a pipeline fashion to reduce the tuning overhead. Three types of multimedia traffic streams, including a constant–bit-rate traffic, a variable-bit-rate traffic with large burstiness, and a variable-bit-rate traffic with longer interarrival times, are considered by the proposed MAC protocol, known as the *multimedia wavelength-division multiple access* (M-WDMA) scheme. The M-WDMA protocol consists of three subprotocols. One is the TDM subprotocol, which is an interleaved TDMA MAC protocol. The second is a reservation-based subprotocol, RSV, which controls the access to the data channels by using a

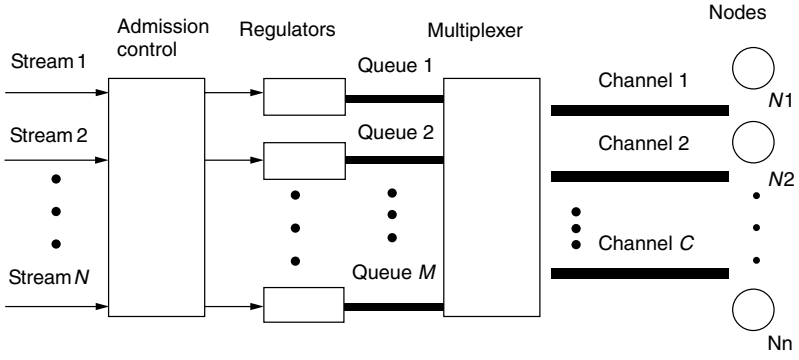


Figure 3. The logic model of the proposed multimedia systematic scheme.

multiple-token method. The third one is a random-access subprotocol, CNT, which works in a way similar to that of the interleaved slotted ALOHA. The outstanding point of this protocol is that a dynamic bandwidth allocation scheme is incorporated into the protocol to dynamically adjust the portions of the bandwidth occupied by the three types of traffic streams according to their QoS demands. The adaptation on the bandwidth allocation can be implemented by adjusting the segment sizes of the timeframe because the different classes of transmissions are grouped into a single timeframe. With knowledge of the size of each segment of a frame, bandwidth allocation can become flexible. A priority scheme is set up to allocate the bandwidth. The TDM requirements have the highest priority, while the requirements of RSV will be considered before the CNT requirements. Using an analytic model and simulation experiments, it has been proved that the performance of the M-WDMA is adequate for WDM optical networks in serving multimedia applications.

4.2. Novel Protocols for Multimedia Applications

A novel MAC protocol for providing guaranteed QoS service to multimedia applications in a WDM optical network [47]. The architecture considered in this network [47] is the same as that in Jia et al. [29]. The MPEG compressed video/audio applications are considered being transmitted in passive star–coupled WDM optical network. The QoS of transmission of the MPEG traffic is based on the frame size traces from the MPEG encoded real video sequences. A frame, which is considered as the basic element with variable size of the MPEG traffic streams, is scheduled and transmitted at one time. A systematic scheme is proposed to guarantee the deterministic delay to the transmission of the MPEG traffic. This scheme includes an admission policy, a traffic characterization mechanism, and a scheduling algorithm as a whole to ensure the QoS of the transmission of MPEG traffic. It is a distributed scheme in the sense that every function of the scheme is implemented at each transmitting node. The logical model of this scheme is shown in Fig. 3.

The admission control scheme is designed to basically limit the number of MPEG traffic entering the network. It is assumed that there are n MPEG traffic sources already connected to the network. There are another m new MPEG traffic sources requesting guaranteed bounded delay service. The admission control scheme employs a transmission bandwidth availability test algorithm to decide whether all the m new traffic sources can be accepted or rejected or which subset of them can be accepted. The traffic characterization scheme simply delays the transmission if the admitted traffic sources do not conform to their characterization to avoid excessive traffic entering into the network. The policy adopted for traffic characterization is based on the concept of the regularity of the marked point process, which is used to model the MPEG traffic. The scheduling algorithm, *adaptive round-robin and earliest available time scheduling* (ARR-EATS), is used as a virtual multiplexer to schedule the transmission of each message in the multiple MPEG traffic sources so that the deterministic delay to each frame of the MPEG streams can be guaranteed.

Analytic evaluation of the guaranteed deterministic delay bound for the proposed system service schemes is based on the theory of max-plus algebra. The deterministic delay bound is verified by intensive trace-driven simulations and modelled MPEG traffic simulations. It is proved that the proposed scheme is efficient and feasible in providing deterministic guaranteed QoS transmission service to the MPEG multimedia applications in the specified WDM optical networks. It is obvious that this protocol stands out as a state-of-the-art MAC protocol among most MAC protocols, which support QoS of the transmission of multimedia applications in WDM optical networks.

An interesting idea regarding the architecture of the WDM optical network has been proposed [48], to support varieties of traffic such as data, real-time traffic, and multicast/broadcast service. The proposed architecture, *hybrid optical network* (HONET), tries to combine the single-hop and multihop WDM optical network architectures into a synergy architecture based on the observation that a multihop network architecture does not efficiently support real-time traffic and multicast/broadcast service because of possible delay fluctuation. A single-hop network is not suitable for providing packet-switched service. The architecture of the HONET can be considered as a network that consists of the multihop network with an arbitrary virtual topology and a single-hop network based on a dynamically assigned T/WDMA MAC protocol. The virtual structured HONET can provide more flexibility in satisfying different application's traffic demands by allowing different network configurations. In this virtual network architecture, real-time traffic and other connection-oriented applications can

be supported by a single-hop network, while non-real-time data traffic, which can tolerate relatively large delay, is supported by a multihop network. The advantage of this virtual architecture is that it is flexible, capable of employing different topologies of the multihop network and different MAC protocols for the single-hop network to support varieties of traffic in the optical network according to the QoS of the traffic demands.

5. SUMMARY

This article has summarized state-of-the-art medium access control protocols for wavelength-division multiplexing (WDM) networks, especially for the passive star–coupled WDM optical networks. Depending on the characteristics, complexity, and capabilities of these MAC protocols, we have classified them as data and message transmission MAC protocols, MAC protocols for real-time transmission service, and MAC protocols for multimedia applications. Most of these protocols focus on local- and metropolitan-area environments. Architectural, qualitative, and quantitative descriptions of various protocols within each category have been provided. Some important or milestone protocols have been given quite detailed explanations to present their underlined significance. It is explicit that real-time message transmission with QoS demands and transmission service to multimedia applications with QoS requirements are currently needed to be supported by the MAC protocols on the WDM optical networks. This article can be used as a good starting point for researchers working on this area to give them an overview of the research efforts conducted since 1990. In addition, the article presents the fundamentals for further investigation into ways of coping with the current and the anticipated explosion of multimedia information transfer.

BIOGRAPHIES

Mounir Hamdi received the B.S. degree in computer engineering (with distinction) from the University of Louisiana in 1985, and the M.S. and Ph.D. degrees in electrical engineering from the University of Pittsburgh in 1987 and 1991, respectively. Since 1991, has been a faculty member in the Department of Computer Science at the Hong Kong University of Science and Technology, where he is now Associate Professor of Computer Science and the Director of the Computer Engineering Programme. In 1999 and 2000, he held visiting professor positions at Stanford University and the Swiss Federal Institute of Technology.

His general areas of research are in networking and parallel computing, in which he has published more than 130 research publications, and for which he has been awarded more than 10 research grants. He has graduated more than 10 MS and PhD students in the area of study. Currently, he is working on high-speed networks including the design, analysis, scheduling, and management of high-speed switches/routers, wavelengeh division multiplexing (WDM) networks/switches, and wireless networks.

Dr. Hamdi has been on the editorial board of *IEEE Transactions on Communications, IEEE Communication Magazine, Computer Networks, Wireless Communication* and *Mobile Computing*, and *Parallel Computing*, and has been on the program committees of mor than 50 international conferences and workshops. He was a guest editor of *IEEE Communications Magazine* and *Informatica*. He received the best paper award at the International Conference on Information and Networking in 1998 out of 152 papers. He received the best 10 lecturers award and the distinguished teaching award from the Hong Kong University of Science and Technology. He is a member of IEEE and ACM.

Maode Ma received the B.S. degree in automatic control in 1982 from Tsinghua University, Beijing, China; the M.s. degree in computer engineering in 1991 from Tianjin University, Tianjin, China; and the Ph.D. degree in computer science from Hong Kong University of Science and Technology in 1999. He joined the computer industry in 1982 as an engineer. In 1986, he was a system enginer at Tianjin University. Starting from 1991, he was an assistant professor in the department of computer engineering at Tianjin University. Since 2000, Dr. Ma has joined the school of electrical and electronic engineering at Nanyang Technological University in Singapore as an assistant professor. Dr. Ma has published approximately 20 academic papers in the areas of WDM optical networks. His areas of research interest are performance analysis of computer networks, optical networks, and wireless networks.

BIBLIOGRAPHY

1. B. Mukherjee, WDM-based local lightwave networks — Part I: Single-hop systems, *IEEE Network* 12–27 (May 1992).
2. B. Mukherjee, WDM-based local lightwave networks — Part II: Multi-hop systems, *IEEE Network* 20–32 (July 1992).
3. I. Chlamtac and A. Ganz, Channel allocation protocols in frequency-time controlled high speed networks, *IEEE Trans. Commun.* **36**(4): 430–440 (April 1988).
4. G. N. Rouskas and M. H. Ammar, Analysis and optimization of transmission schedules for single-hop WDM networks, *IEEE/ACM Trans. Network.* **3**(2): 211–221 (April 1995).
5. A. Ganz and Y. Gao, Time-wavelength assignment algorithms for high performance WDM star based systems, *IEEE Trans. Commun.* **42**(2–4): 1827–1836 (Feb.–April 1994).
6. G. R. Pieris and G. H. Sasaki, Scheduling transmissions in WDM broadcast-and-select networks, *IEEE/ACM Trans. Network.* **2**(2): 105–110 (April 1994).
7. M. S. Borella and B. Mukherjee, Efficient scheduling of nonuniform packet traffic in a WDM/TDM local lightwave network with arbitrary transceiver tuning latencies, *IEEE J. Select. Areas Commun.* **14**(6): 923–934 (June 1996).
8. M. Azizoglu, R. A. Barry, and A. Mokhtar, Impact of tuning delay on the performance of bandwidth-limited optical broadcast networks with uniform traffic, *IEEE J. Select. Areas Commun.* **14**(6): 935–944 (June 1996).
9. P. W. Dowd, Random access protocols for high speed interprocessor communication based on an optical passive star topology, *IEEE/OSA J. Lightwave Technol.* **9**(6): 799–808 (June 1991).

10. A. Ganz and Z. Koren, WDM passive star protocols and performance analysis, *Proc. IEEE INFOCOM'91*, April 1991, pp. 991–1000.

11. I. M. I. Habbab, M. Kavehrad, and C.-E. W. Sundberg, Protocols for very high speed optical fiber local area networks using a passive star topology, *IEEE/OSA J. Lightwave Technol.* **5**(12): 1782–1794 (Dec. 1987).

12. N. Mehravari, Performance and protocol improvements for very high-speed optical fiber local area networks using a passive star topology, *IEEE/OSA J. Lightwave Technol.* **8**(4): 520–530 (April 1990).

13. G. N. M. Sudhakar, M. Kavehrad, and N. Georganas, Slotted ALOHA and reservation ALOHA protocols for very high-speed optical fiber local area networks using passive star topology, *IEEE/OSA J. Lightwave Technol.* **9**(10): 1411–1422 (Oct. 1991).

14. G. N. M. Sudhakar, N. Georganas, and M. Kavehrad, Multicontrol channel for very high-speed optical fiber local area networks and their interconnections using passive star topology, *Proc. IEEE GLOBECOM'91*, Dec. 1991, pp. 624–628.

15. F. Jia and B. Mukherjee, The receiver collision avoidance (RCA) protocol for a single-hop lightwave network, *IEEE/OSA J. Lightwave Technol.* **11**(5–6): 1052–1065 (May/June 1993).

16. M.-S. Chen, N. R. Dono, and R. Ramaswami, A media access protocol for packet-switched wavelength division multiaccess metropolitan area networks, *IEEE J. Select. Areas Commun.* **8**(8): 1048–1057 (Aug. 1990).

17. I. Chlamtac and A. Fumagalli, Quadro-stars: High performance optical WDM star networks, *Proc. IEEE GLOBECOM'91*, Dec. 1991, pp. 1224–1229.

18. M. Chen and T.-S. Yum, A conflictfree protocol for optical WDM networks, *IEEE GLOBECOM'91*, Dec. 1991, pp. 1276–1291.

19. R. Chipalkatti, Z. Zhang, and A. S. Acampora, High-speed communication protocols for optical star networks using WDM, *Proc. IEEE INFOCOM'92*, May 1992, pp. 2124–2133.

20. R. Chipalkatti, Z. Zhang, and A. S. Acampora, Protocols for optical star-coupler network using WDM: Performance and complexity study, *IEEE J. Select. Areas Commun.* **11**(4): 579–589 (May 1993).

21. P. A. Humblet, R. Ramaswami, and K. N. Sivarajan, An efficient communication protocol for high-speed packet-switched multichannel networks, *IEEE J. Select. Areas Commun.* **11**(4): 568–578 (May 1993).

22. H. Jeon and C. Un, Contention-based reservation protocols in multiwavelength optical networks with a passive star topology, *Proc. IEEE ICC*, June 1992, pp. 1473–1477.

23. J. H. Lee and C. K. Un, Dynamic scheduling protocol for variable-sized messages in a WDM-based local network, *IEEE/OSA J. Lightwave Technol.* **14**(7): 1595–1600 (July 1996).

24. K. Bogineni and P. W. Dowd, A collisionless media access protocol for high speed communication in optically interconnected parallel computers, *Proc. SPIE* **1577**: 276–287 (Sept. 1991).

25. P. W. Dowd and K. Bogineni, Simulation analysis of a collisionless multiple access protocol for a wavelength division multiplexed star-coupled configuration, *Proc. 25th Annual Simulation Symp.* April 1992.

26. K. Bogineni and P. W. Dowd, A collisionless multiple access protocol for a wavelength division multiplexed star-coupled configuration: Architecture and performance analysis, *IEEE/OSA J. Lightwave Technol.* **10**(11): 1688–1699 (Nov. 1992).

27. K. M. Sivalingam and P. W. Dowd, A multilevel WDM access protocol for an optically interconnected multiprocessor system, *IEEE/OSA J. Lightwave Technol.* **13**(11): 2152–2167 (Nov. 1995).

28. K. M. Sivalingam and P. W. Dowd, A lightweight media access protocol for a WDM-based distributed shared memory system, *Proc. IEEE INFOCOM'96*, 1996, pp. 946–953.

29. F. Jia, B. Mukherjee, and J. Iness, Scheduling variable-length messages in a single-hop multichannel local lightwave network, *IEEE/ACM Trans. Network.* **3**(4): 477–487 (Aug. 1995).

30. A. Muir and J. J. Garcia-Luna-Aceves, Distributed queue packet scheduling algorithms for WDM-based networks, *Proc. IEEE INFOCOM'96*, 1996, pp. 938–945.

31. B. Hamidzadeh, Maode Ma, and M. Hamdi, Message sequencing techniques for on-line scheduling in WDM networks, *IEEE/OSA J. Lightwave Technol.* **17**(8): 1309–1319 (Aug. 1999).

32. M. Ma, B. Hamidzadeh, and M. Hamdi, A receiver-oriented message scheduling algorithm for WDM lightwave networks, *Comput. Networks* **31**(20): 2139–2152 (Sept. 1999).

33. M. Jonsson, K. Borjesson, and M. Legardt, Dynamic time-deterministic traffic in a fiber-optic WDM star network, *Proc. 9th Euromicro Workshop on Real Time Systems*, June 1997, pp. 25–33.

34. M. Ma, B. Hamidzadeh, and M. Hamdi, Efficient scheduling algorithms for real-time service on WDM optical networks, *Photon. Network Commun.* **1**(2): (July 1999).

35. M. Ma and M. Hamdi, An adaptive scheduling algorithm for differentiated service on WDM optical networks, *IEEE GLOBECOM'01*, 2001.

36. H.-Y. Tyan, J. C. Hou, B. Wang, and C. Han, On supporting temporal quality of service in WDMA-based star-coupled optical networks, *IEEE Trans. Comput.* **50**(3): 197–214 (March 2001).

37. A. C. Kam, K.-Y. Siu, R. A. Barry, and E. A. Swanson, A cell switching WDM broadcast LAN with bandwidth guarantee and fair access, *IEEE/OSA J. Lightwave Technol.* **16**(12): 2265–2280 (Dec. 1998).

38. M. Ma and M. Hamdi, Providing deterministic quality-of-service guarantees on WDM optical networks, *IEEE J. Select. Areas Commun.* **18**(10): 2072–2083 (Oct. 2000).

39. A. Yan, A. Ganz, and C. M. Krishna, A distributed adaptive protocol providing real-time services on WDM-based LANs, *IEEE/OSA J. Lightwave Technol.* **14**(6): 1245–1254 (June 1996).

40. B. Li and Y. Qin, Traffic scheduling in a photonic packet switching system with QoS guarantee, *IEEE/OSA J. Lightwave Technol.* **16**(12): 2281–2295 (Dec. 1998).

41. S. Selvakennedy, A. K. Ramani, M. Y. M-Saman, and V. Prakash, Dynamic scheduling scheme for handling traffic multiplicity in wavelength division multiplexed optical networks, *Proc. 8th Int. Conf. Computer Communications and Networks*, 1999, pp. 344–349.

42. M. A. Marsan et al., All-optical WDM multi-rings with differentiated QoS, *IEEE Commun. Mag.* 58–66 (Feb. 1999).

43. J. Indulska and J. Richards, A comparative simulation study of protocols for a bus WDM architecture, *Proc. Int. Conf. Networks*, 1995, pp. 251–255.

44. W. M. Moh et al., The support of optical network protocols for multimedia ATM traffic, *Proc. Int. Con. Networks*, 1995, pp. 1–5.

45. N.-F. Huang and H.-I. Liu, Wavelength division multiplexing-based video-on-demand systems, *IEEE/OSA J. Lightwave Technol.* **17**(2): 155–164 (Feb. 1999).

46. L. Wang and M. Hamdi, Efficient protocols for multimedia streams on WDMA networks, *Proc. 12th Int. Conf. Information Networking*, 1998, pp. 241–246.

47. M. Ma and M. Hamdi, Providing guaranteed deterministic performance service to multimedia applications on WDM optical networks, *Proc. IEEE GLOBECOM'00*, 2000, Vol. 2, pp. 1171–1175.

48. M. Kovacevic and M. Gerla, HONET: An integrated services wavelength division optical network, *Proc. IEEE ICC'94*, 1994, pp. 1669–1674.

MULTIMEDIA NETWORKING

ANDREA BIANCO
Politecnico di Torino
Torino (Turin), Italy

STEFANO GIORDANO
University of Pisa
Pisa, Italy

ALFIO LOMBARDO
University of Catania
Catania, Italy

1. INTRODUCTION

Multimedia networks are an evolution of integrated networks: networks designed to support different services, each of which is provided through the exchange of a single medium. The term *multimedia* itself denotes, in fact, the integrated manipulation of different media related to a single end-user application environment. The most widely used applications in multimedia networks are data transfer between computers, video/audiostreaming, interactive telephone applications, and voice and videoconferencing. Multimedia networking therefore deals with mechanisms to support real-time and non-real-time media over digital networks.

Distributed multimedia applications have several requirements with respect to the service that are offered by the communication network. These requirements can be classified as *functional* and *traffic requirements* [1].

Functional requirements are related mainly to the support of distributed cooperative services and therefore refer to either multicast transmission or the ability to define coordinated sets of unicast transmissions. Multicast support in the network is fundamental to provide the transmission of a single copy of data to multiple receivers with the purpose of reducing the network load and the processing load at the sender.

Traffic requirements are related to the user-perceived quality of service (QoS) and therefore refer to parameters such as bandwidth, delay, and reliability (or loss probability). *Bandwidth* specifies how much data are to be transferred over a given time period. *Reliability* pertains to the loss and corruption of data. *Delay* determines the temporal relationship between data transmission and reception.

Bandwidth requirements are fundamental for most applications, although some applications, such as data transfer, may be quite tolerant to even highly variable bandwidth. Loss requirements are stringent for most applications, with the exception of uncompressed voice and video applications, which may tolerate some losses. Delay is the main constraint for real-time interactive media, due to users' delay expectations and synchronization requirements. Users' delay expectations are due mainly due to the ability to interact in real time and enforce constraints on the end-to-end average delay. Synchronization is the preservation of time constraints within a multimedia stream at the time of playout. A multimedia stream is made up of multiple monomedia datastreams related to each other by proper timing relationships; for this reason, multimedia applications require preservation of both intramedia synchronization, to maintain the temporal order in each media evolution, and intermedia synchronization, to maintain the temporal order of correlated events in the application. The former is affected by delay jitter introduced in the network; the latter is affected by *skew*, that is, the difference between the delay jitter suffered at a given time t by two synchronized media.

Note that intermedia synchronization usually involves an increase in the QoS requirements of the low constraining media as well; for instance, in a slide show session, where time relationships exist between voice and image, still-picture transmission becomes delay-sensitive because it is related to the audio evolution, whereas in a monomedia stream still-picture transmission is seldom delay-sensitive.

All the QoS parameters mentioned above are closely related — the greater the overall bandwidth required over a link compared to the link capacity, the more messages will be accumulated and the larger the buffer needed to avoid losses; the more the buffer grows, the larger the delay experienced. Moreover, different user applications often require the enforcement of different and contrasting QoS parameters. Consider, for example, a voice phone application that is sensitive to end-to-end delay and quite tolerant to losses, and a data transfer application that, on the contrary, is interested in small losses and tolerant to delay variations. Clearly, it may be difficult to satisfy both requirements at the same time when the two applications share a link, as is the case in multimedia networks. Using large buffers, for example, makes it easier to control losses but typically increases delay. Thus, the design of flexible solutions to efficiently provide a compromise between different and contrasting traffic requirements is the main challenge for multimedia networking today. To satisfy the QoS requirements of different multimedia applications, suitable traffic control and resource management strategies have to be implemented in multimedia networks.

In this article, we will first discuss the possible approaches and key algorithms to efficiently support QoS requirements in a multimedia network. Then, an historical perspective of multimedia networks evolution will be presented. Finally, we will outline the challenging multimedia network paradigms currently being defined by the IETF to support QoS in the Internet arena.

2. QOS PROVISIONING APPROACHES

A multimedia network must be aware of the characteristics of the user traffic in order to manage different traffic flows and provide the required QoS parameters. For this reason traffic sources are usually classified as constant-bit-rate (CBR) sources and variable-bit-rate (VBR) sources. CBR traffic sources are described by means of their bandwidth requirements, as, for example, in the case of a 64 kbps (kilobits per second) digital telephone call. CBR traffic sources are exactly predictable given their bandwidth requirements, since the same amount of data is generated at fixed time instants. Conversely, VBR traffic sources, such as those generated by data transfer applications, are unpredictable because the number of packets and their emission time depend on terminal workload, software implementation and network access protocols. For this reason, VBR sources are described by a set of parameters that have to be representative of the traffic statistics. The most popular set of parameters are the average bandwidth, that is, the bit rate averaged over the flow duration, and peak bandwidth — the maximum bit rate over the flow duration. CBR traffic can be managed more easily by the network, because in this case deterministic rules may be used to assign network resources to the CBR source; this is the approach used, for example, in circuit-switched networks such as ISDN or telephone networks. Unfortunately, even CBR sources, when compressed to reduce their bit rate, become unpredictable sources of traffic. Thus, multimedia networks must deal primarily with VBR sources.

A first approach to supporting VBR traffic with the requested loss and delay requirements would be to overdesign network resources based on VBR source peak bandwidth demand; by so doing, in fact, VBR traffic is managed as CBR traffic simply ignoring VBR traffic fluctuations. Such an approach, often named *overprovisioning*, can be efficient in terms of resource utilization only if VBR traffic is characterized by a low ratio between peak bandwidth and average bandwidth, usually referred to as "burstiness"; if this is not the case, this approach results in low resource utilization and, therefore, high costs. The overprovisioning approach is used today to design access networks based on LAN technologies where no explicit mechanism to support QoS is provided. Several researchers believe that bandwidth costs will drop dramatically in the near future, thanks to the huge amount of bandwidth available on optical fibers. If this is the case, the overprovisioning approach might turn out to be the best one, since it does not require any effort in telecommunication network engineering.

If, on the contrary, killer applications that may cause network collapse are expected to always exist, then resource utilization is very important and resource management and allocation techniques have to be implemented in network nodes to jointly provide QoS and obtain high resource utilization. This approach requires knowledge of user traffic statistics on the basis of known parameters (traffic specification), knowledge of the path or the network portion where the flow will be routed, and implementation of the following key algorithms for traffic control and resources management:

- *Call admission control* (CAC) algorithms implement rules for accepting or refusing a user call (flow) according to the declared traffic parameters and the availability of the resources needed to meet the requested QoS without disrupting the QoS provided to already accepted calls.
- *Shaping* algorithms are used at network edge to make the user traffic compliant with the traffic specifications.
- *Traffic verification* or *policing* algorithms are used at the access node to ensure that the user traffic is compliant with the parameters declared by the source in the traffic specifications.
- *Resource allocation* or reservation algorithms implement functions for allocating, in all the nodes crossed by the traffic flows, the bandwidth and storage resources needed to provide the requested QoS. This allocation must consider resource sharing as a fundamental issue to obtain high network utilization. Bandwidth allocation is usually provided by *scheduling* algorithms that implement functions in each node to transmit at a given time the most urgent or important packet of the available packets so as to satisfy the application needs. *Buffer management* policies are the most important storage allocation techniques.
- *QoS routing* algorithms implement rules for QoS based routing. Unlike the minimum distance or hop count routing strategies used in traditional networks, QoS routing may ease the task of network dimensioning and provide higher network utilization.

An alternative or additional approach may be to introduce *scaling mechanisms* in the multimedia application to dynamically modify the characteristics of the transmitted data stream with the purpose of adapting the workload generated to the resources available in the network. These mechanisms usually use feedback information on the network congestion status as, for example, in the TCP protocol. Of course, the cost of scaling is a decrease in the user perceived quality; moreover, it is not easy to achieve a match between the workload and the available network resources.

In the following sections, we will first outline the main functions of traffic control algorithms and of resources management algorithms. Then scaling techniques for workload adaptation will be introduced.

2.1. Call Admission Control

Call admission results from a negotiation between the user and the network. For this aim the multimedia traffic source is requested to declare both a set of parameters

which characterize its traffic at the network ingress and the QoS it requires.

The task of the CAC algorithm is to determine whether there are enough resources to meet the new call's QoS requirements without violating the QoS provided to already accepted calls. The CAC algorithms run in each node selected by the routing algorithm to support the new call, and they are therefore related to routing algorithms. If enough resources exist in all the nodes, the call is accepted and the data transfer can start; otherwise the call is dropped.

Several CAC algorithms have been proposed in the literature. We can identify two broad families of possible approaches. The first is based on determining an "equivalent bandwidth" [2], that is, a bandwidth required to satisfy the call QoS needs given the traffic characterization. Only if a bandwidth greater than or equal to the equivalent bandwidth is available over each link of the path, will the call be accepted; in this case the equivalent bandwidth is reserved for the call in each node and it is subtracted from the available bandwidth of each link. This approach is very simple and efficient once the equivalent bandwidth is computed. To this end several sophisticated techniques have been proposed. In general, they are based on simulative or analytic paradigms modeling network node behavior, and compute the loss probability and delay experienced when a new call is multiplexed over the node output link together with other calls. The resulting equivalent bandwidth is typically greater than the mean rate of the call, and the more stringent the QoS parameters the closer the bandwidth is to the peak rate. Unfortunately the equivalent bandwidth approach works only if the traffic generated by the user is similar to the traffic model used for equivalent bandwidth computation. Moreover, the complexity of the computation may reduce the practical applicability of this approach.

The second approach is based on network measurements [3] and therefore does not need to assume any specific model for the source or for aggregation of sources. The available bandwidth is measured over each link and a call is accepted if the available bandwidth is greater than or equal to, for example, the peak rate of the call. Once the call has been accepted, the new bandwidth availability for further new calls will be obtained by successive measurements.

In this case CAC algorithms are usually simple since they have to compare a rate that can be very simply derived from the call traffic parameters with the measured available bandwidth over the link. Unfortunately, it is not easy to determine an effective measurement of the available bandwidth. A possible solution to this problem comes from some queuing theory results that link the effective bandwidth to queuing performances; by using these results, it is possible to propose some estimates of the effective bandwidth.

2.2. Traffic Shaping and Policing

A positive result of the CAC procedure commits the multimedia source to guaranteeing that the traffic profile emitted conforms to the declared traffic parameters. In this perspective, the effectiveness of both the shaping functions in the multimedia application and the policing functions performed at network edge constitutes a challenge to meet the requested QoS.

The set of parameters used to characterize user traffic at the network ingress is specific to each network technology. However, the key parameters that must be provided, regardless of network technology, are the average bit rate for CBR traffic, and the average and peak rate (or burstiness) for VBR traffic. It may be useful, and is often required, to provide information relating to the burst duration, that is, the time for which a given source may send data at the peak rate.

Shaping is usually achieved by means of buffers designed not to exceed a given maximum delay at the transmission side. In the case of real time data, particularly voice and video, in order to avoid or decrease the delay introduced, a feedback on the encoding parameters can be used in conjunction with a virtual buffer [4] that, without introducing a delay, can be used to monitor the source emission and to force the encoding process to maintain the parameters declared by the source.

The declared traffic parameters are usually policed by the network at its access point and sometime at network boundaries when data traffic crosses edges between two different network providers. The average rate and the burst duration are usually policed by means of a token bucket device [5]; the peak rate is policed by means of a controller that monitors the interarrival time of the incoming packets. The "token bucket" is a token pool in which the tokens are generated at a constant rate equal to the average rate declared by the multimedia source. The bucket size represents the maximum capacity of the pool and is related to the declared burst duration. When a packet arrives, a number of tokens equal to the packet dimension in bytes is drawn from the pool; if a packet arrives when the pool is empty, it is marked as nonconforming to the traffic specification declared by the multimedia source, and may be dropped.

2.3. Scheduling

Scheduling algorithms are run in nodes to support different levels of priority or urgency criteria for data belonging to different calls. They span from the very simple FIFO (first-in first-out) mechanisms, where data are sent over each output link in the order in which they were received, to strict priority mechanisms, where lower-priority data are sent if and only if no higher-priority data exist in the node, up to sophisticated scheduling algorithms such as weighted round robin (WRR) or weighted fair queuing (WFQ), which are able to provide each call with bandwidth guarantees [6].

Several aspects must be taken into account when considering scheduling algorithms: complexity, the ability to separate flow behavior (a property often referred to as "isolation"), the ability to provide bandwidth guarantees, and buffer sharing capability. It must be noted that buffer management techniques, and in particular the queue architecture adopted within nodes, are strictly connected to the scheduling algorithm. For the sake of conciseness, in this article we focus on traditional output queuing (OQ) node architectures. In this architecture packets arriving

at input links are immediately transferred and stored in buffers at output links; they provide the best performances but require very high speed in the internal switching fabric.

In traditional data networks, nodes usually employ a FIFO [sometimes called first-come first-served (FCFS)] scheduling technique, with a single buffer for each output link shared evenly among all calls. Data are transmitted on output links in the temporal order in which they were received at input links. When the queue becomes full, packets are dropped, regardless of the flow to which they belong, until a position in the queue becomes available. The FIFO scheduling does not provide any form of isolation among flows; all the flows obtain the same QoS, which depends on the behavior of the other flows. Moreover, no bandwidth guarantees can be provided. However, buffers are shared among all flows, and thus fully utilized, and the algorithm is very simple. This scheduling technique may be suited for monomedia networks, where all the flows are interested in the same QoS parameters, but not for multimedia networks.

The key assumption required to provide isolation among flows is to create several queues at each output link and assign each flow to a queue; this is also referred to as *queue partitioning*. If this is the case, all flows belonging to the same queue share the same QoS, whereas isolation among flows assigned to different queues is quite simple to obtain.

In priority-based scheduling algorithms, each queue is associated with a different level of priority. Arriving packets are stored in the queue with the same level of priority as flow to which they belong. Queues are served in strictly increasing order of priority. Packets are extracted from the highest-priority queue until it is empty; and only if this queue is empty does the scheduling look at the next highest-priority queue and so on for all defined priorities. No preemption is enforced on packets being transmitted; thus a higher-priority packet must wait until the transmission of the ongoing lower priority packet has ended. This scheduling is fairly simple and provides isolation among flows belonging to different level of priority. Unfortunately, no isolation exists among flows with the same priority; lower-priority flows may be starved by higher-priority ones and therefore bandwidth is not guaranteed to all flows.

The simplest bandwidth guaranteeing scheduling is the WRR scheduling [7]. In WRR, each queue is assigned a weight, typically related to flow bandwidth needs.[1] The scheduling defines a service cycle; during this cycle, flows are served a number of times proportional to flow weights. Although this idea can provide bandwidth guarantees and flow isolation quite simply, it has several drawbacks. First, the service cycle length depends on the ratio between the flows' bandwidth requests. If we imagine for the sake of simplicity that flows send fixed-size data, the length of the cycle is the MCD of the flow weights. This number can become fairly large. Moreover, calls may be closed when the scheduler is in the middle of its service cycle. In that case a new cycle should be defined, but the part of the previous cycle not served must also be taken into account to meet flow bandwidth requirements correctly. The same holds when a new call is accepted. Several implementations of WRR-like schedulers, based on counters associated to each flow to face the problem of WRR service cycle, have been proposed in the literature [see, e.g., an article on deficit round robin (DRR) [8]].

More complex algorithms derive from the generalized processor sharing (GPS) scheduler [9]. The GPS scheduler emulates the behavior of an ideal fluid system, where each flow is continuously served at a rate proportional to its fair bandwidth share, determined on the basis of the number of active flows and on their required bandwidth needs. The emulation must take into account the constraint that in the real system only complete packets can be sent and flows cannot be served continuously; this give rise to packet GPS (PGPS) scheduling. PGPS schedulers can be implemented by assigning tags to packets when they arrive at input links (tags are roughly inversely proportional to flow rates) and serving packets in order of increasing tags.

Other implementations that approximate the behavior of the ideal fluid system described above have been proposed in the literature starting from the weighted fair queuing (WFQ) scheduling algorithm [10–14]. The most important properties of WFQ schedulers are flow isolation and bandwidth guarantees and, most importantly, if flows are leaky-bucket-controlled and all the nodes implement WFQ scheduling, then bounded end-to-end delay can be computed and guaranteed, as demonstrated in [9].

It must be observed that many other schedulers have been proposed (e.g., EDD, jitter EDD, stop and go; see [6] for a summary they exhibit interesting properties, although for the moment they have been not extensively implemented.

2.4. Buffer Management

Buffer management techniques have always been used, even in traditional network architecture; they become a key issue for multimedia networks, however. The most important buffer management techniques exploit *buffer occupancy* measures, *buffer allocation and partitioning*, and *dropping techniques*.

Buffer occupancy may be used as status information by CAC procedures to assess the network capability to accept calls and is used as a congestion indication by network nodes, providing users with explicit congestion signals.

Buffer allocation and partitioning is a key element in a scheduling technique, as described above, to protect flows from the behavior of other flows. Moreover, in order to provide different flows with bandwidth guarantees, some authors propose controlling only buffer allocation in nodes, since the bandwidth provided on each link is clearly proportional to the number of packets stored for the relative flow.

Packet dropping techniques are fundamental because nodes have to deal with finite buffer capacity. Some dropping techniques have the goal of dropping an entire packet; in networks where packet fragmentation often

[1] To provide bandwidth guarantees, it is necessary either for each flow to declare its bandwidth needs (these may be computed by CAC algorithms as seen before) or for all the flows to agree to share the bandwidth fairly.

occurs as a result of small-size data units, nodes may try to drop fragments belonging to the same packet rather than different packets, given that a single fragment loss renders the remaining portion of the packet useless for the user (data loss or retransmission may occur depending on the application) [15]. Other dropping techniques, based on random choice, try to reduce either correlation among packet losses or synchronization in packet losses among calls. Reducing correlation is useful, since most protocols that deal with packet losses use retransmission techniques that are negatively affected by correlated losses. Reduction of packet loss synchronization among calls can be achieved by spreading packet losses for different calls over time; this improves the performance of TCP, the most widely used transport protocol: since TCP senders reduce the data transmission rate when experiencing losses to prevent network congestion, if too many TCP flows experience losses at the same time, network utilization may drop to very low values even when this is not really necessary. These techniques are generally known as active queue management (AQM) techniques and have received a great deal of attention. As an example, random early detection (RED) [16] and random exponential marking (REM) [17] have been proposed for the Internet arena.

2.5. QoS Routing

Routing is also a fundamental task in multimedia networks. Routing in a topology implies choosing a "best" path (sequence of edges) to connect two nodes according to some defined metric associated with edges (links), and then distributing network information if the link metric dynamically changes with time. Two routing aspects are peculiar to multimedia networks: (1) dealing with QoS requirements when selecting the best path for each flow and (2) providing support for multicast traffic.

QoS routing implies selecting network routes with sufficient resources to satisfy the requested QoS parameters while achieving high resource utilization. The difficulty lies in the fact that QoS routing implies satisfying multiple constraints (e.g., bandwidth, delay, and loss requirements); it is well known that finding a feasible path with even only two independent path constraints is NP-hard. Moreover, multiple constraints impose multiple metrics for each link, thus increasing the difficulty in gathering up-to-date information on network status.

Three broad classes of algorithms proposed for QoS routing exist [18,19]: source routing algorithms, distributed routing, and hierarchical routing algorithms. They can be used together to obtain better performance depending on the network architecture. In *source routing*, each source node keeps information about the network global state (topology and state information on each link) and computes a feasible path locally; each emitted packet contains, therefore, routing information that forces the forwarding action in each node along the path. In *distributed routing*, the path is computed using a distributed algorithm; control messages are exchanged among nodes to create network state information and search for a feasible path. In *hierarchical routing*, nodes are clustered into groups in a multilevel hierarchy. Each node maintains an aggregated global state (partial information) that contains detailed information about nodes in the same group and aggregated information about other groups.

Multicast routing implies finding the best tree connecting a source node with a set of destination nodes [20]. Finding either a least-cost tree or the least-cost tree with bounded delay are both NP-hard problems. Several heuristics have been proposed to solve the multicast routing problem. It is, however, important to point out that QoS routing and QoS multicast routing are definitely the less developed of the techniques to control multimedia traffic described in this article.

2.6. Scaling Mechanisms

Scaling mechanisms have been used up to now to prevent occurrence of congestion in TCP/IP-based networks [21]. They are also used today to provide users with CBR real-time applications even when the compression algorithms used intrinsically lead to VBR streaming. This is the case, for example, of the MPEG video encoder currently used for video transmission in circuit switched network environment such as the ISDN.

Since the late 1990s the use of scaling mechanisms has been proposed for VBR real-time data transmission over packet-switched networks even when no stringent QoS guarantees are provided [22]. Scaling mechanisms can, in fact, be used to adapt the traffic profile emitted by the VBR source according to the variable-bandwidth profile available in a multimedia network.

In a real-time VBR video source, for example, this can be achieved either by changing the video coding parameters runtime or by using hierarchical coding schemes, or again by lowering the frame rate.

To scale the source traffic efficiently, a feedback control loop is introduced to monitor the network status; by so doing, when changes occur in the available network bandwidth, the appropriate actions are taken in the source to change its throughput accordingly. All the control mechanisms proposed use the delay and loss experienced in the end-to-end transmission to compute the available bandwidth, that is, the amount of data that can be transmitted without incurring in QoS degradation.

The protocols defined for this purpose have the target of calculating the bandwidth that TCP congestion control mechanisms make available to the TCP data source, in each network condition. They can be classified into three groups:

- *Window-based congestion control protocols*, which use a congestion window; therefore all the congestion control mechanisms implemented in the different TCP versions belong to this group [21].
- *TCP-like congestion control protocols*, which adapt the transmission rate according to the additive increase multiplicative decrease (AIMD) strategy without using congestion windows [23].
- *Equation-based congestion control protocols*, which adapt the transmission rate according to a suitable equation which, usually, derives from a TCP throughput model [24].

3. MULTIMEDIA NETWORKS: HISTORICAL PERSPECTIVE

Since the 1970s, the challenge of a universal network capable of providing the transport of information related to different media has stimulated the development of several network architectures each characterized by a specific information transfer strategy [25]. The oldest and most widely diffused network architecture is, of course, the telephone network; its technical approach is more than a century old and at the current stages of evolution it supports integrated transmission and switching technologies [Integrated Digital Network (IDN)] and integrated access modalities for services provision (ISDN or narrowband ISDN). This architecture, based on circuit switching, surely represents the best solution for the provision of the QoS required by any end-to-end narrowband communication among CBR applications; it does not, however, possess the flexibility needed to support VBR applications, such as data transfer among computers, since the circuit-switching approach provides only deterministic resource allocation.

In order to support VBR data communication without any performance guarantees (best-effort data communication), the IP datagram approach soon became the de facto standard in all the enterprise LANs, and later on, in LAN and WAN interconnection. This approach, in fact, is based on packet switching techniques that are highly suitable for VBR traffic, thanks to their ability to provide statistical resource sharing. Unfortunately, the connectionless data transfer supported by the IP approach is unable to satisfy the specific QoS needs of different applications.

The design of broadband multimedia network architectures started from these two network realities: a ubiquitous circuit-switched telephone network and a widely diffused packet-switched datagram-based internetwork. In the 1990s in particular the standardization environment that defined the ISDN proposed broadband ISDN and the related asynchronous transfer mode (ATM) technology to support multimedia traffic and its QoS requirements. This choice is based on the virtual circuit approach, which represents a compromise between statistical packet switching based on datagram transmission and deterministic circuit switching based on the synchronous transmission of small fixed-size data units. However, the B-ISDN dream crashed for two main reasons:

- The difficulties of constructing applications directly on top of ATM APIs that require the handling of very powerful, but not user-friendly, ATM control and management facilities
- The need to interoperate seamlessly with LANs such as Ethernet for the provision of efficient IP packet transmission over the ATM.

So, ATM never reached the desktop, and it was relegated to the backbone, whereas the growing success of Ethernet from shared media to switching paradigm lead to *overprovisioning* as the solution for QoS provision in the local and corporate environment.

Most companies today seem to prefer high-speed switched LANs instead of an ATM infrastructure, and IP is clearly the most widely accepted network paradigm today, although the network does not guarantee the correct delivery of packets and may be subject to unpredictable delays. The key question today is therefore how to evolve IP networks toward a multimedia network providing worldwide support for multimedia traffic. This target requires that control and management functions are correctly designed and injected in IP technology.

4. NEXT-GENERATION INTERNET: A CHALLENGING APPROACH FOR MULTIMEDIA NETWORKING

The IETF (Internet Engineering Task Force) has defined two alternative frameworks to provide QoS in a datagram network matching the flexibility required by multimedia networking [26,27]. The first one, referred to as *Internet Integrated Service architecture*, can be considered as a relevant upgrade of the control plane of the classical IP network. New signaling protocols, such as Resource ReserVation Protocol (RSVP), has been introduced to distribute information about resources to be allocated within the network. Unlike ATM or ISDN networks where the signaling protocols produce "hard states", RSVP packets produce "soft state" within the devices (QoS-aware IP router), that is, proper configuration of resources that are automatically released if the "soft state" is not refreshed periodically.

Furthermore, the RSVP signaling paradigm was conceived from the beginning to take into account a multicast environment. For this reason, in fact, in RSVP it is the destination of the multimedia flows that claims for the resources to be allocated to the flow coming from the source; the reservation messages that flow from the destination to the source allow the nodes of a multicast tree to merge resources reservations related to the same multipoint session.

It is very relevant to notice that RSVP is just a signaling protocol able to transfer a request for proper resource allocation among nodes (including the end systems). To be effective, an IntServ network has to implement the functional components introduced in the previous sections in each of its routers. IntServ defines three service classes: the best-effort class, the controlled load service class, and the guaranteed service class.

The first is the present behavior of the Internet (no guarantees), whereas the last one corresponds to detailed (hard) guarantees for bandwidth, delay, and loss. The second class has not been defined in a rigorous manner but is related to behavior that is strictly better than best-effort. It corresponds to the best-effort performances that could be obtained on the network if it were lightly loaded (although these controlled load flows are passing through a network that is highly loaded). It is more a relative behavior that a service class, that is defined on the basis of specific performance parameters (as for the guaranteed service class).

Let us stress here that QoS is provided on a per flow basis in the IntServ architecture, that is

- Every flow is identifiable by means of a specific flow identifier or by other classifications such as the tu-pla consisting of IP-source, IP-dest, TCP-Port-num-source, TCP-Port-num-dest.

- Each end-to-end session in either a unicast or a multicast environment will be composed by several unidirectional flows that will be managed by the networks nodes with per flow queuing, per flow scheduling, shaping, discarding, and so on.

The traffic specification that makes it possible to accept or refuse each flow is very simple and based on a linear upper bound termed the linearly bounded arrival process (LBAP).

Per flow signaling and queuing, of course, produce an impressive increase in the complexity of the network due not only to the RSVP signaling burden on the network but in particular the processing power that is necessary in the router to manage the signaling components and the scheduling schemes. A backbone Internet router, at the time of writing, is loaded by more than 100,000 flows, and although not all of them will need resource reservations, the state and processing requirements within the nodes are prohibitive for scalability reasons.

Because of its scalability problem, IntServ could be considered a suitable architecture for QoS provisioning in IP networks only when there are no more than a 1000 reserved flows. This may be the case of a corporate or campus network, but within the backbone a new approach is needed.

The Internet Engineering Task Force faced the scalability problem of IntServ by defining the *Internet Differentiated Services* architecture, DiffServ for short. The basic idea of Diffserv is to consider an aggregate of flows (called *macroflows*) instead of single flows in an IP QoS domain. The complexity of the nodes is reduced since there is no need to manage the single flows, which lose their identity in an aggregation of many flows with similar requirements. Of course, this is questionable if the different flows have different requirements. Furthermore, the DiffServ architecture defines a core/edge approach were the most complicated activities are carried out by a border router (at the ingress point of the domain) while the core routers operate on the packets in a simpler way. At the ingress of the network the flows are classified as belonging to a certain macroflow by marking them with a marker (code) that is written in the header of the IP packet. To this end, the *type of service* field within the IPv4 header, renamed *differentiated services code point* (DSCP), is used. This field produces a different *per hop behavior* (PHB) in the router. The IETF defined three different possible PHBs: best effort, assured forwarding PHB, and expedited forwarding PHB. Again, the first one is the behavior of the present Internet architecture while the last corresponds to real-time traffic that requires hard guarantees on delay, jitter, loss, and bandwidth.

Assured forwarding is an intermediate class that is further divided into subclasses with different priorities and discarding privileges. Different network domains could associate their specific PHB with the same macroflow, and the relations between different domains have to be negotiated through a proper service-level agreement (SLA) which will initially be almost static and in the future could be set up in a dynamic fashion using proper observation of the state of each domain. For this purpose a possible approach could be based on the adoption of centralized devices called *bandwidth brokers*.

The main limitations of DiffServ are related to the uncertain provisioning of the end-to-end QoS. Associated with a macroflow, in fact, it is not certain whether a single flow will receive the correct performances, particularly if several domains are passed. DiffServ is certainly an oversimplification of the IntServ approach and tries as far as possible to avoid the requirement for signaling from end systems. However, a signaling scheme will definitely be necessary between the nodes and the bandwidth broker.

Whatever approach is used, multimedia will lead to significant problems with respect to the traffic forwarding on the network. *Forwarding* is the process of moving one packet from a certain input to a certain output, and this is done taking into account the information that is passed by the routing process. It is relevant to mention here that the classical routing process is carried out on the Internet via a "destination-based approach." The forwarding process, therefore, is quite complicated if a classless routing scheme is adopted and a "longest matching"[2] lookup has to be carried out for every packet. For this reason a fixed-size approach would be better. This was one of the reasons for the deployment of multiprotocol label switching (MPSL). At the entrance of the network a label is associated with each packet by a classification procedure. Similar to the label swapping techniques adopted in ATM and frame relay networks, this forwarding identification label simplifies the forwarding of variable-size packets to their destination. Furthermore, the path that a packet has to follow on the network can be established by the egress node (i.e., the node at the entrance of the network) and a label distribution procedure can update the forwarding tables, which can now be based on labels instead of the destination address field. In this way it is easy to provide load balancing between the different routes or build up recovery procedure for the traffic that was critically stopped by a fault.

MPLS can distribute and allocate these labels using different approaches, namely, Label Distribution Protocol (LDP), Constraint-based Routing using Label Distribution Protocol (CR-LDP), or RSVP. Again RSVP is used as a signaling protocols extending its functions to provide traffic control (RSVP–traffic Engineering).

The MPLS labels can be stacked (push operation) to funnel some traffic over specific paths and then bring back the identity of each macroflow at the egress nodes (pop operation). This helps in building up virtual private networks (VPNs) at the layer 3 level (by using MPLS it is also possible to build up level 2 VPNs).

It is relevant to mention that MPLS is a network control plane that is unable to provide QoS by itself. The constraint-based routing schemes (such as CSPF) adopted by MPLS can lead to a proper path on the network that takes into account the specific requirements of certain end-to-end sessions, but again it is scheduling, active queue management and shaping algorithms that are really

[2] Longest matching lookup aims to find the network address which best matches the destination written in the header of a packet, in a routing table.

responsible for the different forwarding behaviors offered by a node from its input to its output.

Because of its marking on entry approach, MPLS couples well with a DiffServ architecture, and this is, at the time of writing, the most promising architecture for a flexible, scalable, controllable, and manageable multimedia network.

BIOGRAPHIES

Andrea Bianco is Associate Professor at the Dipartimento di Elettronica of Politecnico di Torino, in Italy. He was born in Torino (Turin), Italy, in 1962. He holds a Dr. Ing. degree in Electronics Engineering (1986) and a Ph.D. in Telecommunications Engineering (1993), both from Politecnico di Torino. From 1987 to 1990 he was with S.S.B. in Torino, where he has been working on office automation projects based on database and distributed networking programming. Since 1994 he was an Assistant Professor at Politecnico di Torino, first in the Dipartimento di Sistemi di Produzione, and later in the Dipartimento di Elettronica. In 1993 he visited Hewlett-Pakard Labs in Palo Alto, California. In the summer 1998 he visited the Electronics Department at Stanford University, California. He has co-authored over 80 papers published in international journals and presented in leading international conferences in the area of telecommunication networks. His current research interests are in the fields of protocols for all-optical networks and switch architectures for high-speed networks. A. Bianco is a member of IEEE.

Stefano Giordano received the Laurea degree "cum laude" in Electronics Engineering from the University of Pisa in 1990 and the Ph.D. degree in Information Engineering in 1994. During 1988/89 he worked with CNR-CNUCE. Since year 1991 he was with the University of Pisa participating and coordinating several research activities sponsored (among others) by Saritel, Siemens, Italtel, CNR, RAI, HP, ASI, TILab, Marconi, and Finsiel. Since 2001 he has been an Associate Professor at the Department of Information Engineering of the University of Pisa, where he give lectures on telecommunication networks and design and simulation of telecommunication networks. His research and professional areas of interest are broadband multimedia communications and telecommunication networks analysis and design. He is responsible for the NetGroup at Consorzio Pisa Ricerche and the TLC NetLab of the Faculty of Engineering in Pisa. Stefano Giodano is participating to the Technical Committee of the Campus Network of the University of Pisa (SERRA), has been a member of the IEEE Communication Society since 1989, of the Internet Society since its foundation, and of the IFIP Working Group 6.3. He was referee of the projects of the European Union, the National Science Foundation, and the Ministry of Research in Italy in the area of telecommunications.

Alfio Lombardo received his degree in electrical engineering from the University of Catania, Italy, in 1983. Until 1987, he acted as consultant at CREI, the center of the Politecnico di Milano for research on computer networks, where he was involved in European research projects on protocol design (SEDOS and CTS-WAN projects). There he was the Technical Coordinator of the Formal Description Techniques (FDT) COST 11 TER project from 1986 to 1988. In 1988 he joined the University of Catania, where he is Full Professor of Telematics. There he was the leader of the University of Catania team in the European ACTS project DOLMEN (Service Machine Development for an Open Long-term Mobile and Fixed Network Environment). Presently, he is involved in the European IST Project VESPER (Virtual Home Environment for Service Personalization and Roaming Users) as leader of the University of Catania team. His research interests include distributed multimedia applications, multimedia traffic modeling and analysis, Internet2, and wireless networks. His email address is *lombardo@iit.unict.it*.

BIBLIOGRAPHY

1. L. C. Wolf, C. Griwodz, and R. Steinmets, Multimedia communication, *Proc. IEEE* **85**(12): (Dec. 1997).
2. R. Guerin, H. Ahmadi, and M. Naghshineh, Equivalent capacity and its application to bandwidth allocation in high-speed networks, *IEEE J. Select. Areas Commun.* **9**(7): (Sept. 1991).
3. S. Jamin, P. Danzig, S. Shenker, and L. Zhang, A measurement-based admission control algorithm for integrated service packet networks, *IEEE/ACM Trans. Network.* **5**(1): (Feb. 1997).
4. A. Lombardo, F. Cocimano, A. Cernuto, and G. Schembra, A queueing system model for the design of feedback laws in rate-controlled MPEG video encoders, *Trans. Circuits Syst. Video Technol.* (in press).
5. C. Partridge, *Gigabit Networking*, Addison-Wesley, Reading, MA, 1994.
6. H. Zhang, Service disciplines for guaranteed performance service in packet-switching networks, *IEEE Proc.* **83**(10): (Oct. 1995).
7. M. Katevenis, S. Sidiropoulos, and C. Courcoubetis, Weighted round-robin cell multiplexing in a general-purpose ATM switch chip, *IEEE J. Select. Areas Commun.* **9**(8): (Oct. 1991).
8. M. Shreedhar and G. Varghese, Efficient fair queuing using deficit round-robin, *IEEE/ACM Trans. Network.* **4**(3): (June 1996).
9. A. K. Parekh and R. Gallager, A generalized processor sharing approach to flow control in integrated services networks: The single-node case, *IEEE/ACM Trans. Network.* **1**(3): (June 1993).
10. A. Demers, S. Keshav, and S. Shenkar, Analysis and simulation of a fair queueing algorithm, *Internet Res. Exp.* **1**: (1990).
11. D. Varma and D. Stiliadis, Hardware implementation of fair queueing algorithms for asynchronous transfer mode networks, *IEEE Commun. Mag.* (Nov. 1997).
12. L. Zhang, Virtual clock: A new traffic control algorithm for packet switching networks, *ACM SIGCOMM'90*, Philadelphia, Sept. 1990.
13. J. C. Bennet and H. Zhang, WF^2Q: Worst-case fair weighted fair queueing, *INFOCOM'96*, March 1996.
14. J. C. Bennet and H. Zhang, Hierarchical packet fair queueing algorithms, *IEEE/ACM Trans. Network.* **5**(5): (Oct. 1997).

15. A. Romanow and S. Floyd, Dynamics of TCP traffic over ATM networks, *IEEE J. Select. Areas Commun.* **13**(4): (May 1995).

16. S. Floyd and Van Jacobson, Random early detection gateways for congestion avoidance, *IEEE/ACM Trans. Network.* (Aug. 1993).

17. S. Athuraliya, S. Low, V. Li, and Y. Qinghe, REM: Active queue management, *IEEE Network* **15**(3): (May–June 2001).

18. H. Chen and K. Nahrstedt, An overview of quality-of-service routing for the next generation high-speed networks: Problems and solutions, *IEEE Network Mag.* (*Special Issue on Transmission and Distribution of Digital Video*) **12**(6): (Nov.–Dec. 1998).

19. E. Crawley, R. Nair, B. Rajagopalau, and H. Sandick, A Framework for QoS-based routing in the Internet, Internet RFC 2386 (April 1998).

20. C. Diot, W. Dabbous, and J. Crowcroft, Multipoint communication: A survey of protocols, functions, and mechanisms, *IEEE J. Select. Areas Commun.* (April 1997).

21. R. Stevens, *TCP/IP Illustrated*, Addison-Wesley, Reading, MA, 1994.

22. S. Floyd and K. Frdl, (Promoting the use of end-to-end congestion control in the Internet), unpublished, Feb. 1998; *http://www-nrg.ee.lbl.gov/floyd/papers.htmlJend2end-paper.html*.

23. R. Rejaie, M. Handley, and D. Estrin, RAP: An end-to-end rate-based congestion control mechanism for realtime streams in the Internet, *Proc. INFOCOM'99*, New York, (March 1999).

24. S. Floyd, M. Handley, J. Padhye, and J. Widmer, *TCP-Friendly Rate Control* (*TRFC*)*: Protocol Specification*, IETF, (July 2001).

25. M. Decina and V. Trecordi, Convergence of telecommunications and computing to networking models for integrated services and applications, *Proc. IEEE* **85**(12): (Dec. 1997).

26. Z. Wang, *Internet QoS Architectures and Mechanism for Quality of Service*, Morgan Kaufmann, 2001.

27. B. Teitelbaum, Internet2 Qbone: Building a testbed for differentiated services, *IEEE Network* **13**(5): 8–16 (Sept.–Oct. 1999).

MULTIMEDIA OVER DIGITAL SUBSCRIBER LINES

HAITAO ZHENG
Bell Laboratories
Lucent Technologies
Holmdel, New Jersey

K. J. RAY LIU
University of Maryland
College Park, Maryland

1. INTRODUCTION

Multimedia communications has become one of the fastest-growing and yet most challenging fields for both academia and industry. Internet-enabled applications such as videoconferencing, multimedia mail, video-on-demand, HDTV broadcast — either by wirelines such as asymmetric digital subscriber lines (ADSLs), integrated services digital networks (ISDNs), or by wireless networks — present new problems of distinctive nature. The high volume of multimedia data can be handled efficiently only if all system resources are carefully optimized. The distinctive nature of multimedia data also requires existing transmission systems to be augmented with functions that can handle more than ordinary data.

Many advances in multimedia communications are made through interaction and collaboration between multimedia source coding and channel optimization. In this article, we present an approach to providing reliable and resource efficient multimedia services through ADSL by jointly considering compression/coding and channel optimization techniques.

We begin this article with a brief review of multimedia communications in general, focusing on multimedia compression/coding and joint source channel optimization. We then describe the channel characteristics and modulation procedure for ADSL. We then examine two transmission architectures for delivering multimedia content over ADSL. The important concept of resource allocation is discussed. The remainder of the article deals with technical details on underlying techniques for these two architectures to perform source/channel optimization and resource allocation. The performance is examined in some practical applications.

1.1. Multimedia Compression and Layered Coding

Multimedia communications benefits greatly from developments in source compression algorithms and hardware implementations. The objective of compression is to reduce the amount of data necessary to reproduce the original data while maintaining a desired level of signal quality and implementation complexity. Compression is necessary and important to reduce the bit rate for efficient transmission over normally bandwidth-limited networks. For digital data that cannot afford to lose any information, the compression schemes used are mainly lossless; that is, the reverse procedure can reproduce the exact original signal. Multimedia data, however, are subject to human perception. For image, video, speech, and audio, loss of some fidelity is tolerable as long as it is not perceivable. Therefore, compression may discard some information in order to achieve more compactness, which corresponds to the well-known lossy scheme.

In the past, the design of multimedia source coding was based mostly on the assumption of error free channels. The objective of compression was to reduce the information rate to the maximum extend without huge quality degradation. It was not recognized that compression, however, renders the compressed data highly vulnerable to channel errors and losses. A few bit errors could lead to severely disrupted media. Researchers have now realized the importance of joint consideration of the distortion induced by channel errors into source coding design. Among all the techniques, layered and scalable coding has been widely recognized and utilized since it can provide error resilience necessary for noisy channel transmissions. One major role of layered coding is to classify media signals in terms of importance and separate them into different layers.

The importance is often defined perceptually. The layers are compressed by different coding schemes and protected by different priority levels. The theme is to assign the highest priority to the most important layer. As such, layered coding achieves error resilience by preventing the loss of perceptually important information. Another advantage of layered coding is scalability. Data rate has a direct impact on transmission cost and media quality. If required, layered coding can alter the data rate to compromise any change in transmission cost and media quality requirements.

Some popular forms of layered coding are subband/wavelet coding and scalability options in H.263+, H.263++, and MPEG-4 [1–5]. It is worth pointing out that integrated services also display a level of scalability. Internet applications are often formulated as a mixture of data, speech and video services, each associated with different data rates and QoS (quality of service) requirements. They can be viewed as a set of layers of different importance. MPEG-4 also specifies audiovideo objects (AVOs) to represent integrated services.

1.2. Joint Source and Channel Optimization

Channel transmission inevitably introduces errors and losses. Multimedia communications systems, including both source and channel coding, have to be robust against channel errors so that media application will not be seriously disrupted by channel errors. In other words, multimedia communications requires efficient and powerful error control techniques.

In general, there are two approaches for error control and recovery. The first approach involves channel transmission design to reduce occurrence of channel errors. Powerful channel coding and Automatic Retransmission reQuest (ARQ) are two commonly used methods. However, the amount of channel coding is limited by bandwidth requirements, and the number of retransmissions is limited by delay requirements. Therefore, the channel transmission technique cannot fully remove channel errors. Given this fact, the second approach appends redundancy in compressed data so that channel error effects can be concealed and even become imperceptible. This is the so-called error concealment and recovery [6]. The design of a concealment strategy depends on the whole system design. More redundancy at source compression results in better concealment. However, it also implies increased bandwidth requirements on channel transmission. For a given bandwidth, increasing the source rate results in reduced channel coding gain and vice versa. As such, the optimal solution, namely, a joint source and channel optimization (JSCO) approach, would be to jointly optimize source and channel coding, and balance the amount of robustness between these two functions.

Both literature and practice have shown that joint optimization leads to substantial gain in performance [7]. For example, layered coding produces layers or classes with different error sensitivities. Unequal error protection, which assigns different loss rates to the layers, yields better media quality compared to a single-layer approach [8]. The joint optimization has two aspects: media quality and resource consumption. The ordinary source coding design only measures media quality, whereas the traditional channel optimization only takes into account resource consumption. The joint optimization aims to "minimize resource consumption while maintaining the desired media quality" or "achieve the best media quality for a set of preassigned resources." Optimization involves selecting source coding parameters (source coding rate, scalability format, etc.) and channel transmission parameters (such as transmit power, channel coding, and modulation). In the following discussion, we assume that the two end systems are connected by a direct link, namely, a circuit switch system. For a packet-switched system, the packetization strategy should also be considered in the optimization. Additional details on packet-based multimedia communications can be found in the literature [9,10].

2. FUNDAMENTALS OF ADSL

The exponential growth of Internet traffic has driven the demand for additional bandwidth and propelled the development of many new mechanisms of transmitting information. One of the most efficient mechanisms is digital subscriber lines (DSLs). DSLs can deliver megabit connectivity to the mass households using traditional phone lines. Mostly favored by Internet users, ADSL [11], asymmetric DSL, is specifically designed to support asymmetric data traffics to exploit the one-way nature of most multimedia applications where large amount of information flows toward the subscribers and only a small amount of interactive control information is transmitted in the upstream direction.

It is necessary to first understand the characteristics of the ADSL channel. Major channel impairments are attenuation and crossstack, which tend to increase as a function of frequency and distance. The resulting channel is spectrally shaped, as shown in Fig. 1. The wide variation in frequency leads to considerable difficulty and complexity for channel equalization, which is necessary for any single-carrier system. To avoid this problem, the ADSL

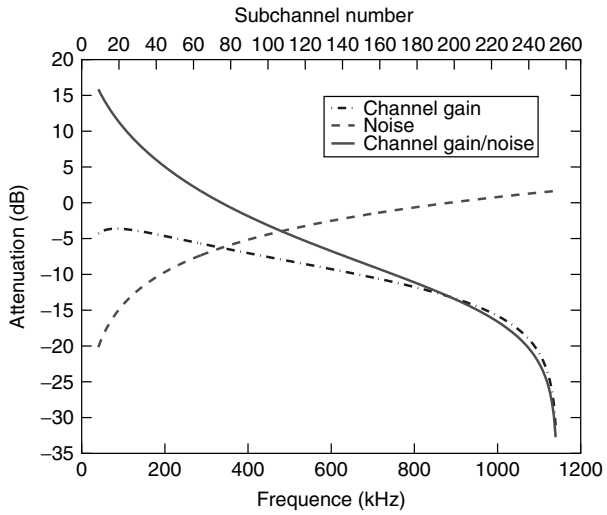


Figure 1. Typical spectrally shaped ADSL channels. The ADSL frequency band spans from 40 to 1397 kHz, with 256 subchannels of 4.3125 kHz each.

community adopted multicarrier modulation (MCM) [12] as the standard channel coding scheme. MCM partitions the ADSL frequency band into a set of independent subchannels, each corresponding to a smaller frequency band. When the number of subchannels is large enough, the subchannels are sufficiently narrow, and can be approximated as independent AWGN channels requiring little or no equalization. Mathematically, each subchannel can be described by

$$R_k = H_k S_k + N_k \tag{1}$$

where S_k and R_k represent the transmitted and the received signal at the kth subchannel and H_k and N_k represent the subchannel (complex-valued) gain and noise, respectively, approximated by the corresponding value at the center frequency of the kth subchannel.

One remarkable merit of MCM is that it allows the data rate, transmit power, and channel coding scheme at each subchannel to change independently. This flexibility provides optimality and fast adaptation to channel variations. As such, the optimal use of the entire channel can be achieved by making optimum use of each subchannel. Associated with subchannels are two transmission parameters: transmit power and data rate. Theoretically, waterfilling can achieve the ultimate system capacity, which allocates different transmit power levels to subchannels with different channel gains. The higher the channel gain, the larger the amount of transmit power. Accordingly, those subchannels with higher power level can transmit at a higher data rate to maximize the overall data rate. There has been extensive study on the allocation of power and data rate to the subchannels, known as the loading algorithm [13–17]. We will review these algorithms in the following sections. More information about ADSL can be found in the book by Starr et al. [18].

3. ARCHITECTURE DESIGN

A primary approach to resource efficiency in multimedia communications is to provide different priorities to layers of different perceptual importance. The architecture design for delivering multimedia over ADSL follows this approach. Although ADSL has the distinct characteristics of a multichannel structure, a rather simple solution — serial transmission — is to view the ADSL channel as a single transmission pipe and design source coding independent of multichannel optimization. Another solution — parallel transmission — combines ADSL channel partitioning and optimization into the system design, which was first proposed for image transmissions in 1998 [19] and extended to video transmissions in 1999 [20].

3.1. Serial Transmission

This approach ignores the multichannel structure within the source coding design. Therefore, joint source channel optimization involves a single transmission pipe and a source coder. Different priority levels can be achieved by transmitting source layers separately in time, and assigning different amount of channel resources. Precisely, the transmission is time-slotted based on where in each time slot only data from one source layer can be transmitted; that is, source layers are time-multiplexed. The optimization allocates time slots to source layers, and within each time slot, distributes channel resources among subchannels. Figure 2a depicts an example of serial transmission with three source layers. Layer 1 is transmitted in time slots 1 and 2, while layers 2 and 3 are transmitted in slots 3 and 4, respectively. Within time slots that a single source layer is transmitted, all the usable subchannels share the same error performance, thus the same priority. However, error performances across time slots in which different source layers are transmitted are completely different. This conclusion is reflected in Fig. 3a, where the error performance is represented by the bit error rate (BER) [20]. It also implies that the system has to optimize resource allocation for each layer. This requirement results in not only additional computational complexity but also extra difficulty for transmitter/receiver implementation due to frequent channel parameter changes. A detailed resource allocation scheme will be discussed in the following section.

3.2. Parallel Transmission

In general, ADSL channels vary slowly in time and may be considered as static. The channel gain and noise variations in the frequency domain are more severe compared to that in time domain. Serial transmission eliminates frequency variations through resource allocation. This variation, however, can be utilized to provide different error priorities to source layers. Such consideration leads to the invention of parallel transmission, which transmits source layers simultaneously in time but at different frequencies. As shown in Fig. 2b, each source layer occupies a certain number of subchannels, corresponding to frequency multiplexing [20]. The optimization requires a subchannel-to-layer assignment, which assigns subchannels to transmit each source layer. It is beneficial to assign subchannels with better channel gain and noise performance to transmit important layers. Such consideration could guarantee reliable base layer transmission with little power consumption — an important advantage of parallel transmission, especially under a low power constraint. Parallel transmission can also integrate traffic flows of various QoS requirements without frequently changing channel settings. Compared to serial transmission, parallel transmission has a completely different error performance distribution. It is observed from Fig. 3b that the error performance varies across a group of subchannels and remains static in time. The technical details of resource allocation will be discussed in next section.

4. SYSTEM OPTIMIZATION

Having described the architecture design for multimedia over ADSL, we now discuss the joint optimization procedure. Depending on the system goal, the optimization problem can be formulated in the two aspects discussed in Sections 4.1 and 4.2.

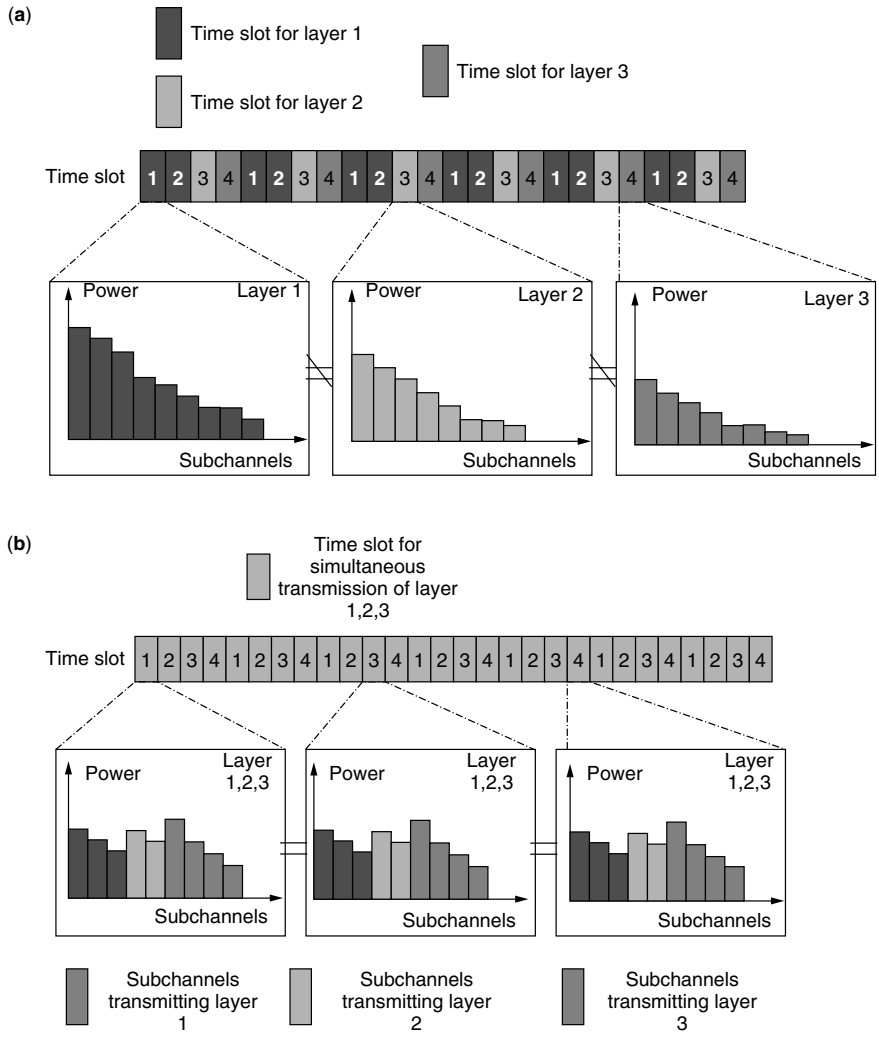


Figure 2. Example architecture for (**a**) serial transmission and (**b**) parallel transmission.

4.1. Optimization 1: Minimizing Cost

The viability of multimedia services business depends on a solution to provide desirable services at a low cost. Cost, within a communication system, is determined mainly by system resource consumption, such as bandwidth, transmit power, and hardware complexity. In this study, we assume fixed bandwidth and interpret power consumption as a major indicator of resource consumption. We intend to find a simple resource allocation scheme due to hardware complexity limitations. The optimization emphasizes subchannel resource allocation and time/subchannel to layer assignment, and intends to minimize transmit power and satisfy QoS requirements. We assume that QoS requirements are represented by the data rate R and the error rate BER.

The optimization consists of two loops. The first loop involves subchannel resource allocation, which is performed by a loading algorithm. On the basis of the QoS requirement, the system allocates transmit power and data rate to subchannels to minimize the overall power consumption. Most existing loading algorithms aim at achieving the same error performance on all the usable subchannels. Given the BER and R, theoretically, the amount of power allocated to subchannel k is proportional to $\Gamma(\text{BER})/g_k$, where g_k represents the channel gain to noise ratio (CGNR) at subchannel k [21]. $\Gamma(\ )$, a function of BER, measures the SNR distance from Shannon capacity. For uncoded QAM, a BER of 10^{-6} corresponds to a Γ of 8.8 dB; while at zero BER, $\Gamma = 10$ dB [18]. Once the amount of transmit power is derived, the corresponding rate at subchannel k can be computed and a modulation is selected accordingly. After loading, the sum of the subchannel transmit power is

$$E(R, \text{BER}, C) = \sum_{k=1}^{C} \frac{\Gamma(\text{BER})}{g_k}(2^{b_k} - 1) \tag{2}$$

where C represents the number of subchannels that are transmitting during this particular layer's transmission; b_k represents the bit rate at subchannel k where $R = \sum_{k=1}^{C_m} b_k$. More details about the loading algorithms can be found in the literature [13–16].

For a given time/subchannel to layer assignment, this inner loop optimization finds the optimal subchannel power and bit rate distribution for each layer. For each source layer, serial transmission performs power and bit

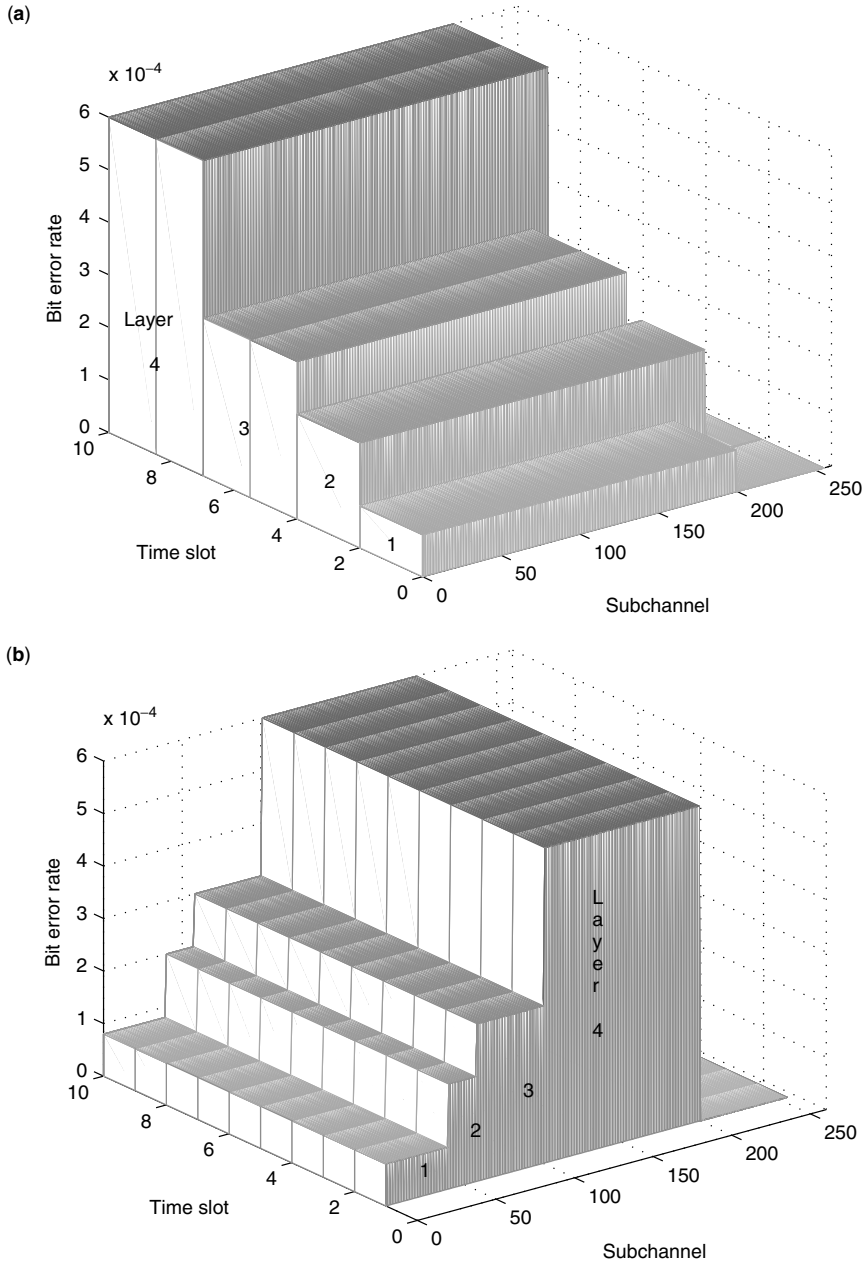


Figure 3. Error performance across the time slots and subchannels for (**a**) serial transmission and (**b**) parallel transmission.

loading using all the subchannels, while parallel transmission performs loading using a group of subchannels. The complexity depends on the number of subchannels and the number of source layers. The outer loop optimization finds the optimal time/subchannel to layer assignment.

- *Serial Transmission: Time-Slot Assignment*. A set of T time slots are grouped together into a time frame, where slot assignment is repeated frame by frame. The optimization can be represented mathematically as follows:

$$\text{Given} \quad \{R_m, \text{BER}_m\}_{m=1}^{N}$$

$$\text{Find} \quad \{T_m\}_{m=1}^{N} \quad \text{where} \quad \sum_{m=1}^{N} T_m = T \quad \text{to} \quad (3)$$

$$\text{Minimize} \quad E_T = \sum_{m=1}^{N} \frac{T_m}{T} E\left(\frac{T}{T_m} R_m, \text{BER}_m, C\right)$$

where T_m represents the number of time slots within a timeframe for layer m, R_m is the throughput requirement of layer m, C represents the number of subchannels, and E_T represents the power constraint. The optimal slot assignment that minimizes the overall transmit power E_T can be solved by an exhaustive search. The complexity depends on the number of slots per frame. A large number of slots results in fine granularity and better performance at the cost of higher complexity. The best solution is a compromise between quality and complexity.

- *Parallel Transmission: Subchannel-to-Layer Assignment.* Obviously, subchannels with higher CGNR

should transmit layers of higher importance. By sorting the subchannels in a decreasing CGNR order, the problem of subchannel-to-layer assignment can be reduced to finding the optimal number of subchannels for each source layer, $\{C_m\}_{m=1}^N$. After sorting, subchannels indexed 1 to C_1 are used to transmit layer 1 while subchannels indexed C_1+1 to C_1+C_2 are for layer 2, and so on. The loading algorithm derives the optimal power and rate distribution for each group of subchannels. Mathematically, the problem is equivalent to

$$\begin{aligned}
\text{Given} \quad & \{R_m, \text{BER}_m\}_{m=1}^N \\
\text{Find} \quad & \{C_m\}_{m=1}^N \quad \text{where} \quad \sum_{m=1}^N C_m \le C \quad \text{to} \\
\text{Minimize} \quad & E_T = \sum_{m=1}^N E(R_{m,T}, \text{BER}_m, C_m) \\
& = \sum_{m=1}^N \sum_{k=C_{m-1}+1}^{C_m} \frac{\Gamma(\text{BER}_m)}{g_k}(2^{b_k}-1) \\
\text{where} \quad & \sum_{k=C_{m-1}+1}^{C_m} b_k = R_m
\end{aligned} \tag{4}$$

Similarly, an exhaustive search can certainly lead to the optimal solution. The complexity depends on the number of source layers and more importantly, the number of subchannels. For ADSL, the number of subchannels is normally more than 256, which implies huge complexity. Using an efficiency measure, a successive search algorithm can quickly approach the optimal solution without examining all the subchannel-layer combinations [22].

4.2. Optimization 2: Quality Optimization

Sometimes, service providers aim to deliver the best service at a fixed cost budget. For most media applications, quality is measured by the distortion between the original and reconstructed data. Mathematically, the distortion can be approximated by the sum of source coding induced distortion D_s and channel transmission induced distortion D_c

$$D = D_s(R_s) + D_c = D_s(R_s) + \sum_{m=1}^N Pe_m W_m \tag{5}$$

where W_m represents the average distortion caused by a single bit error at layer m and Pe_m represents the BER for layer m. The source coding scheme determines the values of R_s, D_s, and $\{W_m\}_{m=1}^N$. When these are given, the goal of minimizing distortion is equivalent to finding the best BER distribution, $\{Pe_m\}_{m=1}^N$ for a given amount of power usage. Thus, BER in this problem, becomes an optimization parameter. Using the loading algorithm defined by Fisher and Huber [17], BER for layer m after loading can be computed as

$$Pe_m(R_m, E_{m,T}, C_m) \approx 4Q\left(\sqrt{\frac{3E_{m,T}G_m/C_m}{(2^{R_m/C_m}-1)}}\right) \tag{6}$$

where

$$G_m = \frac{C_m}{\sum_{i=C_{m-1}+1}^{C_m} 2^{(b_i-R_m)}/g_i} \tag{7}$$

represents the rate averaged CGNR for layer m, and

$$E_{m,T} = \sum_{i=C_{m-1}+1}^{C_m} E_i \tag{8}$$

represents the total power consumption of the subchannels assigned to layer m. We refer $E_{m,T}$ as layer power. $Q(x)$, the Q function, is defined by

$$Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,dt \tag{9}$$

For a given R_s, the optimization problem focuses on D_c only

$$\begin{aligned}
\text{Given throughput} \quad & \{R_{m,T}, W_m\}_{m=1}^N \\
\text{Find} \quad & \{C_m, T_m\}_{m=1}^N \quad \text{to} \\
\text{Minimize} \quad & D_c = \sum_{m=1}^N Pe_m(R_m, E_m, C_m, T_m) W_m \\
\text{subject to} \quad & \sum_{m=1}^N \frac{T_m}{T} E_m \le E_T, \sum_{m=1}^N C_m \le C
\end{aligned} \tag{10}$$

where E_T is the total power constraint and C is the maximum number of subchannels [20].

- *Parallel Transmission*. For a given subchannel to layer assignment $\{C_m\}_{m=1}^N$, the loading algorithm optimizes each layer's power consumption to satisfy a total power contraight:

$$E_{m,T} = \Phi_{\alpha_m}^{-1}\frac{\lambda_{\text{opt}}}{W_m}, \qquad \text{where} \qquad \Phi_\alpha(x) = \sqrt{\frac{\alpha}{x}}\exp(-\alpha x) \tag{11}$$

and λ_{opt} satisfies

$$\sum_{m=1}^N \Phi_{\alpha_m}^{-1}\frac{\lambda_{\text{opt}}}{W_m} = E_T, \quad \text{where} \quad \alpha_m = \frac{3G_m}{2C_m(2^{R_{m,T}/C_m}-1)} \tag{12}$$

To find the optimal $\{C_m\}_{m=1}^N$ to minimize D_c, a successive search can be applied with reasonable computational complexity [20].

- *Serial Transmission*. The time-slot-to-layer assignment $\{T_m\}_{m=1}^N$ requires a power allocation at the source layer level to achieve different error performance during different time slots. We define the power consumption of all the subchannels during layer m's transmission to be e_m. For a given $\{T_m\}_{m=1}^N$, the optimal $\{e_m\}_{m=1}^N$ can be resolved by finding a λ such that [20]

$$E_T = \sum_{m=1}^N \frac{T_m}{T}\Phi_{\beta_m}^{-1}\left(\frac{\lambda}{(1-\rho_m)W_m + \rho_{m+1}W_{m+1}}\right) \tag{13}$$

where

$$e_m = \Phi_{\beta_m}^{-1}\left(\frac{\lambda}{(1-\rho_m)W_m + \rho_{m+1}W_{m+1}}\right)$$

$$\beta_m = \frac{3}{C_m(2^{R_{m,T}/C_m}-1)\dfrac{1}{C_m}\displaystyle\sum_{i=1}^{C_m}\frac{1}{g_i}2^{(R_i-R_{m,T}/C_m)}}$$

$$C_m \le C, m = \cdots N$$

5. SOME APPLICATIONS

Having studied the optimization algorithms for both serial and parallel transmissions, we now present some applications.

5.1. Image and Video

Today's Internet applications such as e-commerce require significant amount of image downloading. Subband/wavelet coding has been a well-known scheme for image compression [1]. The compressed image consists of a set of subbands with different level of perceptual importance. In this example, we consider the quality maximization problem, where image quality is measured by peak signal-to-noise ratio PSNR = 10 log (255^2/MSE), where MSE represents the mean-squared error between the original image and the reconstructed image, a widely used distortion measure. The power constraint is represented by the power usage averaged over the subchannels. Figure 4a depicts the image PSNR as a function of the power constraint E_{av}. A grayscale image "Lena" is subband-coded and quantized to achieve a source data rate of 0.1 and 0.5 bit per pixel (bpp). The ADSL channel consists of 256 subchannels, and supports QAM-64, QAM-32, QAM-16, QAM-8, and QAM-4 modulation types. This example proves that parallel transmission outperforms serial transmission by 4–10 dB in terms of PSNR [20].

Compared to image transmission, video transmission requires more sophisticated optimization because of its large volume, variable data rate, and sensitivity to channel errors. Recent advances in video coding emphasize its error resilience features, mostly in terms of scalability. Using H.263 low bit rate video as an example, low-frequency coefficients, particularly DC coefficients, reflect higher importance compared to high-frequency coefficients; motion vectors have more impact on the decoded video quality than DCT coefficients if corrupted. We use the error-resilient entropy code (EREC) to separate video signals into layers of different importance. EREC is widely recognized since it reorganizes variable-length blocks to fixed-length slots such that each block starts at a known position and the beginning of each block is more immune to error propagation than those at the end [23]. Figure 4b depicts the video quality in terms of averaged PSNR over 60 frames of QCIF (176 × 144 pels) color sequence "Miss America." Similarly, we observe 2–4 dB improvement by parallel transmission compared to serial transmission [20]. The gain is smaller than that of image transmission, since importance classification is much more difficult in video coding.

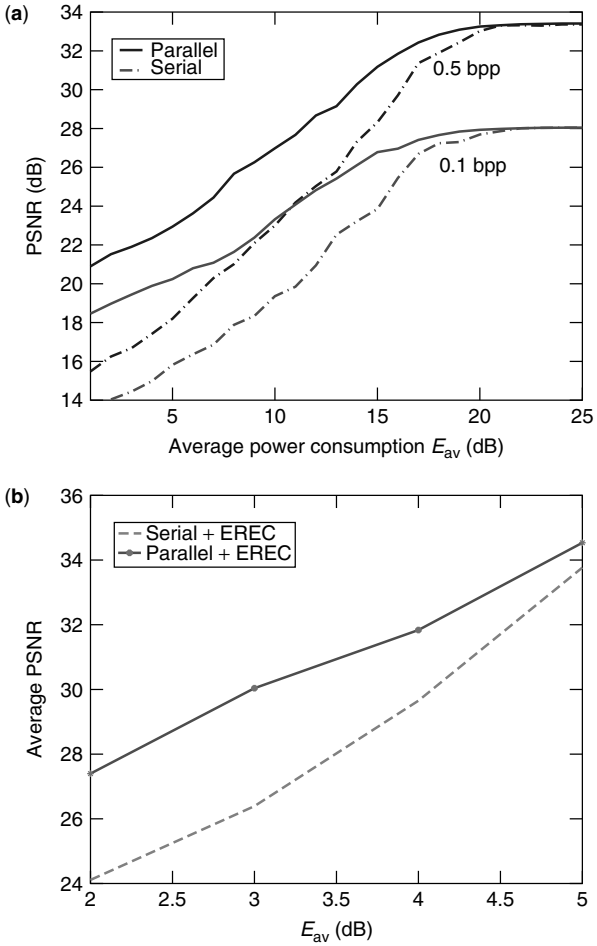


Figure 4. Received PSNR performances for (**a**) image "Lena" and (**b**) video "Miss America."

5.2. Integrated Service of Video, Speech, and Data

Many Internet applications involve integrated services. We select three services of 200, 64, and 10 kbps. We vary the BER distribution to achieve different QoS profiles. Table 1 illustrates the profile characteristics [22]. The outcome of cost/power minimization is shown in Fig. 5 in terms of the transmit power consumption of both serial and parallel transmissions. Similar to the previous examples, parallel transmission outperforms serial transmission by reducing power consumption by 0.5 to 1 dB. To examine the impact of modulation types, we compare the power usage when employing 5 and 11 modulation types. As shown in Fig. 5a, increasing the modulation type can further reduce power usage by 2–3 dB. During power and bit rate loading, the highest modulation type bounds the

Table 1. QoS Requirements

Requirement	Service 1 200 kbps	Service 2 64 kbps	Service 3 10 kbps
QoS1	10^{-6}	10^{-5}	10^{-3}
QoS2	10^{-5}	10^{-3}	10^{-6}
QoS3	10^{-3}	10^{-5}	10^{-6}

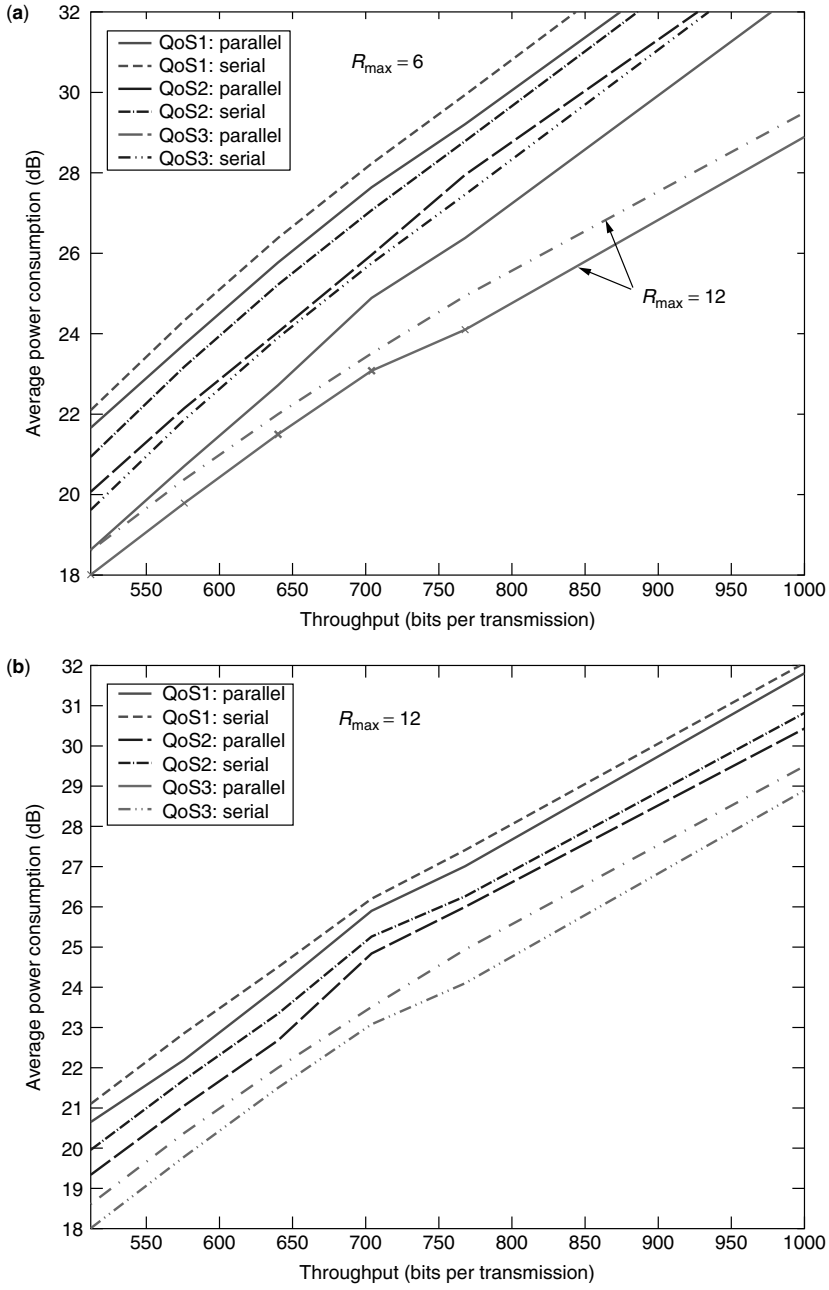


Figure 5. Transmitted Power Consumption vs. Data Throughput for (**a**) $R_{\max} = 6$ and (**b**) $R_{\max} = 12$. The ADSL system is configured to have 256 subchannels and using QAM modulations with $R_{\min} = 2$.

subchannel bit rate. If the allocated bit rate is higher than that offered by the highest modulation, the bit rate is set to that of the highest modulation and the remaining power is allocated to other subchannels. This certainly results in suboptimality. However, the number of modulations also reflects hardware complexity. Quality and complexity should be considered jointly on implementation.

6. CONCLUSION

In this article, we have discussed a key aspect of designing high-performance multimedia communications systems, the ability to perform efficient resource allocation. Most multimedia content consists of layers with different priorities. Unequal error protection, achieved either by source coding or during channel transmission, can effectively reduce resource consumption. We explored the concept of joint optimization in ADSL in terms of two transmission architectures: serial and parallel transmissions. We examined the resource allocation problem in two aspects: cost minimization and quality optimization. The performances of individual allocation algorithms are examined by several practical examples. The study indicates that the parallel transmission architecture, which utilizes the ADSL channel characteristics to provide unequal error protection to source layers, can effectively reduce resource consumption and achieve desirable media quality.

After reading this article, the readers should be able to understand the basic framework of a multimedia communication system, centering on joint source and channel optimization.

BIOGRAPHIES

Haitao Zheng received the B.S. degree in electrical engineering from Xian Jiaotong University, People's Republic of China, in 1995 and the MS and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, Maryland in 1998 and 1999, respectively.

From 1995 to 1998, she was an Institute for System Research Fellow at University of Maryland, College Park. She received the 1998–1999 George Harhalakis Outstanding Systems Engineering graduate Student Award in recognition of outstanding contributions in cross-disciplinary research from the University of Maryland, College Park. Since August 1999, she has been with Wireless Research Laboratory, Bell Labs, Lucent Technologies in Holmdel, New Jersey. Her research interests include design and performance analysis for wireless communications with an emphasis on MAC/PHY layer design, and signal processing techniques for multimedia communications.

K. J. Ray Liu received the B.S. degree from the National Taiwan University and the Ph.D. degree from UCLA, both in electrical engineering. He is a professor in the Electrical and Computer Engineering Department of the University of Maryland, College Park. His research interests span broad aspects of signal processing architectures; multimedia communications and signal processing; wireless communications and networking; information security; and bioinformatics, in which he has published over 230 refereed papers, of which more than 70 arc in archival journals.

Dr. Liu is the recipient of numerous awards, including the 1994 National Science Foundation Young Investigator, the IEEE Signal Processing Society's 1993 Senior Award, and the IEEE 50th Vehicular Technology Conference Best Paper Award, Amsterdam, 1999. He also received the George Corcoran Award in 1994 for outstanding contributions to electrical engineering education and the Outstanding Systems Engineering Faculty Award in 1996 in recognition of outstanding contributions in interdisciplinary research, both from the University of Maryland.

Dr. Liu is editor-in-chief of EURASIP *Journal on Applied Signal Processing* and has been an associate editor of *IEEE Transactions on Signal Processing*; a guest editor of special issues on Multimedia Signal Processing of Proceedings of the IEEE; a guest editor of a special issue on Signal Processing for Wireless Communications of the *IEEE Journal of Selected Areas in Communications*; a guest editor of a special issue on Multimedia Communications over Networks of the *IEEE Signal Processing Magazine*; a guest editor of a special issue on Multimedia over IP of *IEEE Transactions on Multimedia*; and an editor of the *Journal of VLSI Signal Processing Systems*.

BIBLIOGRAPHY

1. J. W. Woods and S. D. O'Neil, Subband coding of images, *IEEE Trans. Acoust. Speech Signal Process.* **34**: 1278–1288 (Oct. 1986).
2. M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
3. T. Gardos, H.263+: The new ITU-T recommendation for video coding at low bit rates, *Proc. IEEE ICASSP*, May 1998, Vol. 6, pp. 3793–3796.
4. J. D. Villasenor and D. S. Park, *Proposed Draft Text for the H.263 Annex V Data Partitioned Slice Mode for Determination at the SG Meeting*, ITU SG16, Proposal Q15 I14, Oct. 1999.
5. *Overall View of the MPEG-4 Standard*, ISO/IEC JTC1/SC29/WG11 N4030, March 2001; *http://mpeg.telecomitalialab.com/standards/mpeg-4/mpeg-4.htm.*
6. Y. Wang, S. Wenger, J. Wen, and A. K. Katsaggelos, Error resilient video coding techniques, *IEEE Signal Process. Mag.* **17**(4): 61–82 (July 2000).
7. S. B. Z. Azami, P. Duhamel, and O. Rioul, Combined source channel coding: Panorama of methods, *Proc. CNES Workshop on Data Compression*, Toulouse, France, Nov. 13–14, 1996.
8. M. Garrett and M. Vetterli, Joint source/channel coding of statistically multiplexed real-time services on packet networks, *IEEE Trans. Network.* **1**(1): 71–79 (Feb. 1993).
9. K. Stuhlmuller, M. Link, and B. Girod, Scalable Internet video streaming with unequal error protection, *Proc. Packet Video Workshop 99*, April 1999, New York.
10. H. Zheng and J. Boyce, An improved UDP protocol for video transmission over Internet-to-wireless networks, *IEEE Trans. Multimedia* **3**(3): 356–365 (Sept. 2001).
11. K. Maxwell, Asymmetric digital subscriber line: Interim technology for the next forty years, *IEEE Commun. Mag.* 100–106 (Oct. 1996).
12. J. A. C. Bingham, Multicarrier modulation for data transmission: An idea whose time has come, *IEEE Commun. Mag.* 5–14 (May 1990).
13. M. Barton and M. L. Honig, Optimization of discrete multitone to maintain spectrum compatibility with other transmission systems on twisted copper pairs, *IEEE J. Select. Areas Commun.* **13**(9): (Dec. 1995).
14. P. S. Chow, J. M. Cioffi, and J. A. C. Bingham, A practical discrete multiton transceiver loading algorithm for data transmission over spectrally shaped channels, *IEEE Trans. Commun.* **43**(2): 773–775 (Feb.–April 1995).
15. B. S. Krongold, K. Ramchandran, and D. L. Jones, Computationally efficient optimal power allocation algorithm for multicarrier communication systems, *Proc. Int. Conf. Communications* (*ICC98*), Atlanta, GA, June 1998.
16. J. Campello de Souza, Optimal discrete bit loading for multicarrier modulation systems, *IEEE Symp. Information Theory*, Boston, 1998.
17. R. F. H. Fisher and J. B. Huber, A new loading algorithm for discrete multiton transmission, *Proc. GlobalCOM 96*, pp. 724–728.
18. T. Starr, J. M. Cioffi, and P. J. Silverman, *Understanding Digital Subscriber Line Technology*, Prentice-Hall, Englewood Cliffs, NJ, 1999.
19. H. Zheng and K. J. R. Liu, A new loading algorithm for image transmission over noisy channel, *Proc. 32nd ASILOMAR Conf. Signal, Systems and Computers*, Pacific Grove, CA, Nov. 1998.
20. H. Zheng and K. J. R. Liu, Robust image and video transmission over spectrally shaped channels using multicarrier

modulation, *IEEE Trans. Multimedia* **1**(1): 88–103 (March 1999).

21. C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* **27**: 379–423 (1948).
22. H. Zheng and K. J. R. Liu, Power minimization for integrated multimedia service over digital subscriber line, *IEEE JSAC* (*Special issue on Error Robust Transmission of Images and Video*) **18**(6): 841–849 (June 2000).
23. D. W. Redmill and N. G. Kingsbury, The EREC: An error-resilient technique for coding variable-length blocks of data, *IEEE Trans. Image Process.* **5**(4): 565–574 (April 1996).

MULTIPLE ANTENNA TRANSCEIVERS FOR WIRELESS COMMUNICATIONS: A CAPACITY PERSPECTIVE

CONSTANTINOS B. PAPADIAS
Global Wireless Systems Research
Bell Laboratories, Lucent Technologies
Holmdel, New Jersey

1. INTRODUCTION

The use of multiple antennas for wireless systems, sometimes referred to as "the spatial frontier," is expected to affect considerably the operation of *wireless networks* [1]. Traditionally, arrays of multiple antenna elements ("antenna arrays") are employed at the base station, due to the associated cost, size, and power constraints that makes their use in wireless terminals more challenging. Used in conjunction with transmit processing on the downlink (base to terminal), or with receive processing on the uplink (terminal to base), *antenna arrays at the base station can offer important performance gains to wireless systems*. By synthesizing spatial beams at the uplink, the base station receiver amplifies the signal-to-noise ratio of a desired user, giving rise to the so-called antenna gain. This gain may in turn be cashed in different ways, such as in an increase of the throughput (data rate), quality, or range of the link. Further, the intelligent shaping of beams (to which the widely used term "smart antennas" is owed) allows one to attenuate the interference that comes from undesired users or undesired locations. An example of such interference mitigation with base station antennas is the use of sectorization in wireless systems; by spatially shaping multiple sectors in each cell, the in-cell interference experienced by each user drops, allowing the co-existence of more users in the cell (higher cell capacities).

On top of the SINR (signal-to-interference-plus-noise ratio) gains described above, antenna arrays may also offer protection against the temporal fluctuations of radio signals, commonly referred to as "fading." *Channel fading is one of the most severe impairments of radio channels*, and its successful handling impacts severely the performance of wireless systems. The best-known way to combat fading is the combining of a number of independently faded received replicas of the transmitted signal. When done cleverly, this (so-called diversity combining) reduces the fluctuation of the signal strength of the received signal against the background noise. As a result, the chance of a deep fade of the received signal is reduced, thus reducing the probability of an outage, specifically, of the situation wherein the link is dropped due to the poor quality of the received signal. As one would expect, the success of diversity combining relies on the degree of statistical independence (typically quantified through the cross-correlation) between the different diversity branches. The spatial dimension, that is, the availability of antenna elements in disjoint spatial locations, is a prime way of obtaining disjointly faded replicas of the transmitted signal for diversity combining. For example, a mobile user located in the proximity of two different base stations may communicate its information to both base antennas. Since the channel is likely to fade in an independent fashion on these two links, the combination of the two independent replicas can offer a substantial diversity gain.

Mechanisms similar to those described above for the realization of performance gains in the uplink can be used in the downlink. For example, if the base station knows the location of a user, it may direct to it a narrow spatial beam. By doing so, it increases the signal-to-noise ratio (SNR) of the user's terminal. Moreover, it helps reduce the interference directed toward users in different locations. To combat fading, the base station may use two antenna elements in order to transmit the downlink signal to a user. For example, by separating widely the two antenna elements, the SNRs of the two links to the user fluctuate in a rather independent fashion. If the base station happens to know at every time instant which of the two links toward the user experiences the best SNR, it may use the corresponding antenna to transmit the signal to the user. This will improve the fading statistics of the signal at the receiver, resulting again in improved reception.

More recently, a number of exciting novel results have given a new push to the field of multiple antennas. This recent revolution began with a research breakthrough by G. J. Foschini at Bell Laboratories in 1996 [2]. Up to that time, the ultimate limit of the spectral efficiency of a wireless link had been studied only for single-antenna systems (or, at most, for single-transmit multiple-receive antenna systems), and it is governed by Shannon's classic capacity formulas for noise-limited channels. In [2] the capacity of wireless links that are equipped with multiple antennas on both sides of the link were studied for the first time; quite astonishingly, the derived capacity formulas showed a spectacular increase in spectral efficiency with the number of transceiver antennas. Roughly speaking, it was shown in [2] that, *when the scattering environment between the multiantenna transmitter and receiver is rich enough, the capacity of the link is roughly proportional to the minimum of the number of antennas on each side* — that is, a doubling of the number of antenna elements on each side of the link is expected to roughly double its spectral efficiency! The capacities predicted by these formulas were unprecedented in the wireless community and have created a large amount of work in order to derive schemes that are capable of delivering them.

In the following, we will describe the main principles that govern the use of multiple antennas in wireless

communication systems. For a compact presentation, we will assume a simple (flat-fading) channel model, which allows us to describe the most important tradeoffs. Moreover, we will focus on link-level studies, for which the metric of Shannon capacity will provide good guidance about both the limitations and success of the described techniques.

2. BACKGROUND AND ASSUMPTIONS

In this section we will outline our assumptions about the considered multiple antenna systems and will provide a corresponding mathematical signal model. We will also briefly review Shannon's classical capacity formula, as it was formulated in the context of single-input/single-output (SISO) systems.

Figure 1 shows a generic architecture of a wireless communication link with M transmitter and N receiver antennas. Such a multiple-input/multiple-output (MIMO) system will be denoted in the remainder of the article as (M, N). As shown in the figure, an original information sequence $\tilde{b}(i)$ that is intended for wireless transmission, undergoes a demultiplexing into multiple data streams before being fed to the transmit antennas. It also typically undergoes forward error correction, interleaving, and spatial multiplexing before being transmitted. Moreover, these operations may happen in a different order. In this article we will be concerned mainly with the channel capacity that can be carried by MIMO channels. However, we will also mention some simple space–time transmission techniques that attempt to attain these capacity bounds.

The demultiplexing/encoding operations result into L data sequences (called *substreams*), denoted by $b_1(k), \ldots, b_L(k)$ (typically $L = M$). After being spatially multiplexed, they are converted into an ensemble of M transmit signals, which are then upconverted to radio frequencies and fed each on every transmit antenna. We will denote the baseband substream transmitted from the mth antenna by $\{s_m(k)\}$. We assume that the physical channel between the mth transmitter and the nth receiver antenna is flat-faded in frequency, so that it can be represented, at baseband, by a complex scalar h_{nm}. The baseband received signal at the receiver antenna array is then represented by a $N \times 1$ vector, denoted by $\mathbf{x}(k)$, that is related to the transmitted substreams as

$$\mathbf{x}(k) = \mathbf{H}\mathbf{s}(k) + \mathbf{n}(k) \tag{1}$$

where $\mathbf{s}(k) = [s_1(k) \cdots s_M(k)]^T : M \times 1$ vector snapshot of transmitted substreams, each assumed of equal variance σ_s^2
$\mathbf{H} = N \times 1$ channel matrix
$\mathbf{x}(k) = N \times 1$ vector of received signal snapshots
$\mathbf{n}(k) = N \times 1$ vector of additive noise samples, assumed i.i.d. and mutually independent, each of variance σ_n^2

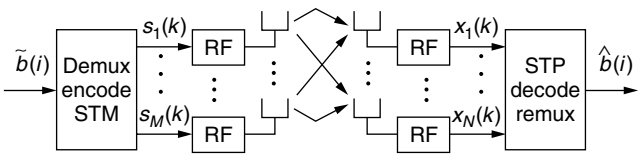


Figure 1. A generic (M, N) multiple antenna transceiver architecture.

We also denote by superscript $*$, T, $\dagger$ the complex conjugate, transpose, and Hermitian transpose, respectively, of a scalar or matrix.

Throughout the remainder of the article, we have chosen to use channel capacity, in the Shannon sense, as a metric for the evaluation of various wireless MIMO systems. Traditionally, smart antennas were viewed as a technology that can help increase the range of a wireless system, the quality of wireless voice calls, or the number of voice users that can be supported in the cell area. However, the advent of MIMO systems since 1996 or so has added a new dimension to the smart antenna technology. It allows us to increase the spectral efficiency of wireless links, without increasing the transmit power; that is, for the same power budget, higher data rates (information throughputs) can be attained.

Channel capacity, in the Shannon sense, is the prime theoretical tool for evaluating the maximum information throughput that can be supported by a communication link. We will hence use it for evaluating the capacity of MIMO systems. Moreover, we will show that, in hindsight, the use of the channel capacity metric offers interesting intuition about the performance of more conventional smart antenna systems. While it can be argued, and rightfully so, that channel capacity is not the sole tool for analyzing wireless systems, we believe that it sheds some light on a number of important issues in smart antenna systems.

As a background to the remainder of the article, we review briefly the channel capacity of single-transmit/single-receive (1, 1) systems. In this case, the (assumed flat-faded) channel is modeled through the complex scalar h, which represents its complex gain. The capacity of this channel, assuming additive white Gaussian noise (AWGN), independent from the transmitted signal, is given by Shannon's capacity formula:

$$C = \log_2(1 + \rho|h|^2) \quad \text{(bps/Hz)} \tag{2}$$

where ρ is the signal-to-noise ratio (SNR) defined as $\rho = \dfrac{\sigma_s^2}{\sigma_n^2}$, σ_s^2 and σ_n^2 being the signal and noise variance, respectively.

Note that the formula indicates that capacity grows logarithmically as a function of the SNR ρ. In other words, any increase in the link's SNR will only be reflected into a corresponding *logarithmic* increase in the channel's capacity. Consequently, each extra bit per second per Hertz (bps/Hz) of capacity, requires roughly a doubling of the link's SNR. This results in a high price for extra capacity — for a linear increase in spectral efficiency, the power needs to be increased exponentially!

3. THE NOTION OF RANDOM CAPACITY

Before discussing the extension of the (1, 1) capacity in Eq. (2) to MIMO systems, we believe that it is important to first underline the random character of wireless systems.

The capacity expression in (2) — similar to the MIMO capacity expressions that follow — is silent about one important issue that underlies the performance of wireless systems in practice, namely, their *statistical behavior*. This is due to the fact that it implicitly assumes the channel to be constant forever, that is, *static*. This point of view leaves out two important aspects of wireless systems:

1. The temporal variation of the channel: each user's (link) changes with time, mainly due to the user's and the environment's mobility
2. The spatial distribution of channels in a geographic area

In other words, in a wireless system, the channel is typically not static, when seen from the perspective of each user, and it is not uniformly distributed at the level of a user population. In one approach that is often taken in order to make capacity expressions relevant to practical systems, the user channel $\mathbf{H}$ is assumed to be *semistatic*. This means that, during time intervals that are of finite duration, but at the same time long enough to allow the desired benefit of error correction coding, the channel is considered static. However, from one such time interval to another, the channel is assumed to be different. *The capacity of such a semistatic channel can then be modeled as a random variable* C, where a pool (statistical ensemble) of channel realizations $\{\mathbf{H}\}$ corresponds to the different time intervals wherein the channel remains static. The capacities corresponding to each realization constitute an ensemble of capacities $\{C\}$ (capacity distribution). The attributes of this ensemble, contained in its cumulative density function (CDF) can be then evaluated in order to assess the statistical capacity of the system. The two most commonly used measures are

1. *Outage capacity* — the capacity point of the CDF that happens with probability higher than a certain target threshold:

$$C_o = \{C \text{ such that } \Pr\{C \geq C_o\} = P_o\} \tag{3}$$

   where P_o is the predetermined outage target. The quantity $1 - P_o$ is often called the *outage probability*. Typically, in cellular voice systems, P_o is chosen to be around 90%, corresponding to an outage probability of 10%. This means that there will be only a 10% probability of the system being in outage, that is, of a user not being able to reach the capacity target C_o.
2. *Average capacity* — the expected value of capacity over the entire CDF:

$$C_a = E(C) \tag{4}$$

   where E denotes statistical expectation. This measure is not traditionally used in voice systems, where it is important to guarantee good (low-latency) service with high probability. However, it appears that it may be a more relevant quantity in wireless data systems, where the bursty nature of data, combined with the higher tolerance to latency and support from higher network layers, makes average throughput a relevant quantity.

The statistical characterization of capacity outlined above does not apply only to time-varying wireless channels. Equally importantly, it can be used to characterize a wireless system in terms of *spatial user distribution*. To illustrate this point, consider a wireless system with static users that are randomly distributed in the geographic area surrounding the base station. Such a scenario applies for example with good accuracy to so-called fixed wireless systems, where the base station communicates with a number of static rooftop antennas. Another example would be that of very low-mobility users getting wireless service in a local-area network (LAN). In such cases, it is important to know what is the capacity as a function of user geographic location. This can be captured in a CDF for the entire ensemble of locations of interest. The notions of outage and average capacities apply here, too. The former conveys the percentage of spatial locations that can be supported with a certain capacity, whereas the latter relates to the total data throughput provided to the entire set of locations. For a service provider, spatial outage may be used as a metric to guarantee a certain quality of service to the customers, whereas the average throughput may relate more to the total revenue per time unit expected in a certain service area.

4. OPEN- AND CLOSED-LOOP MIMO LINK CAPACITIES

In this section, we will present the theoretical maximum spectral efficiencies (channel capacities) that are achievable in multiple antenna links. For a moment, we will neglect the random character of channel capacity; the capacity expressions that will be presented will be subject to specific channel instantiations (channel snapshots). Later on, though, these instantaneous capacity expressions will be numerically evaluated over statistical ensembles that correspond to either channel or user behavior, in order to capture the random aspect of capacity mentioned above.

In our treatment, we will consider the general case of an arbitrary number of antennas on each side of the link. As capacity depends on the knowledge of the MIMO channel response at the transmitter, we will treat separately the two extreme cases: the *open-loop capacity*, which assumes no channel knowledge at the transmitter; and the *closed-loop capacity*, which assumes full channel knowledge at the transmitter. As mentioned above, the presented formulas will be used in subsequent sections in order to analyze special cases and to evaluate the merits of different transmission/reception schemes.

4.1. Open-Loop Capacity

In the open-loop case, the Shannon capacity of the (M, N) flat-faded channel is given [2] by the now familiar (so-called "log-det") formula:

$$C = \log_2 \left\{ \det \left(I_N + \frac{\rho}{M} \mathbf{H}\mathbf{H}^{\dagger} \right) \right\} \quad \text{(bps/Hz)} \tag{5}$$

where the SNR is now defined as $\rho = M\sigma_s^2/\sigma_n^2$. A first important observation that can be made from this equation is that, for rich scattering channels, the MIMO channel capacity grows roughly proportionally to the minimum of the number of transmitter and receiver antennas [2–4].

This is an astonishing result, contrasted to the logarithmic capacity increase as a function of signal-to-noise ratio. Its consequences are quite dramatic as far as the spectral efficiency of a wireless link with a certain power budget is concerned. For example, consider a (1, 1) system that operates at 0 dB SNR. From Eq. (2), its average capacity over an ensemble of i.i.d. Gaussian (Rayleigh-amplitude) channel coefficients h with $E|h|^2 = 1$ is equal to $C_{a,1} = 0.85$ bps/Hz. At the same SNR, a (10, 10) system whose channel coefficients between any transmitter/receiver pair are again independently fading i.i.d. unit-variance Rayleigh variables, has an average capacity of $C_{a,10} = 8.38$ bps/Hz. This corresponds roughly to a 10-fold increase in capacity using 10 antennas on each side of the link and keeping the total transmission power constant ($C_{a,10}/C_{a,1} = 9.76$). Notice that a 10-fold capacity increase in the original (1, 1) system would require to double the power 9 times, that is, increase the power by $2^9 = 512$ times!

Figure 2 shows a graphical depiction of the linear capacity increase with the number of transmit antennas at different SNRs. We have plotted capacities at the 10% outage level for symmetrical MIMO systems that have up to 20 antennas on each side of the link. Each channel realization is a square matrix whose elements are chosen independently from a Rayleigh distribution of unit variance. Notice the extraordinarily high capacities achieved with many antennas (on the order of hundreds of bps/Hz). For comparison, it should be kept in mind that current cellular wireless systems operate typically at no more than 2 bps/Hz. It is also worth noting that, *in MIMO systems, the higher the SNR, the steeper the capacity increase*.

4.2. Closed-Loop Capacity

A generalization of the capacity formula (5) to the case where the transmitter has some knowledge of the channel characteristic $\mathbf{H}$ was derived in [3]:

$$C = \log_2 \left\{ \det \left(I_N + \frac{\rho}{P_T} \mathbf{H} \Phi \mathbf{H}^\dagger \right) \right\} \quad \text{(bps/Hz)} \qquad (6)$$

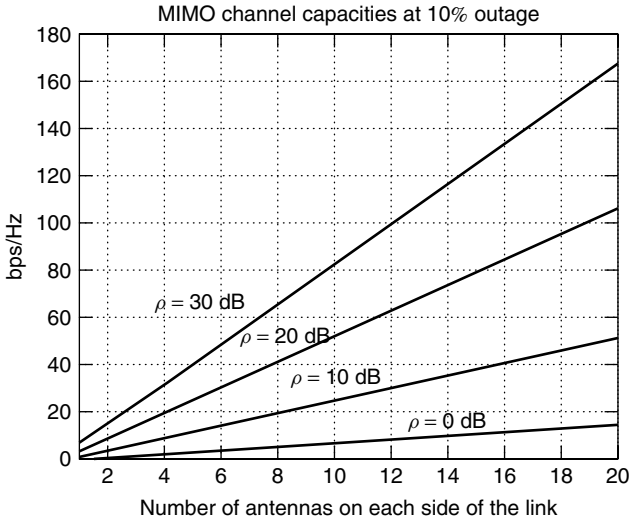


Figure 2. Channel capacity of MIMO systems as a function of the number of antennas.

where Φ is the $M \times M$ covariance matrix of the transmitted signal ($\Phi = E(\mathbf{s}\mathbf{s}^\dagger)$) and P_T is the total transmitted power from the M antennas [$P_T = M\sigma_s^2$, and $tr(\Phi) = P_T$, where $tr()$ denotes the trace of a matrix]. When the channel $\mathbf{H}$ is fully known at the transmitter, Φ in Eq. (6) can be optimized by so-called "spatial water filling." Spatial waterfilling is a notion similar to the one of frequency waterfilling (water pouring) used in orthogonal frequency division multiplexing (OFDM) systems in order to maximize capacity; it is based on the idea of distributing the available transmitted power in a nonuniform fashion across the different "modes" of a channel. In the case of narrowband MIMO systems that we are examining, these are *spatial* modes, defined as the eigendirections of the transmitted signal covariance matrix Φ. The use of a matrix Φ that is not a multiple of an identity matrix denotes a nondiagonal loading of the channel's spatial modes.

In the following, some special cases of (6) will help illustrate the concept of spatial waterfilling. For the moment, note that by choosing $\Phi = (P_T/M)\mathbf{I}_M$, (6) reduces to the open-loop capacity in (5). This represents the fact that when the transmitter has no channel knowledge, diagonal loading of the modes is used, resulting in the open-loop capacity (5).

5. (1, N) SYSTEMS

In this section, we will consider single-input/multiple-output (SIMO) systems. This case is frequently encountered in practice, for example, in the uplink (terminal-to-base) of cellular systems. For reasons of cost, complexity, power, and size, mobile terminals are typically equipped with a single antenna, whereas today's base stations already use two and will soon use even more antennas for reception. We should also note that SIMO systems are, in many ways, more conventional than both MISO (multiple-input/single output) and MIMO (multiple-input/multiple output) systems. This has to do with the fact that, when having a single transmitter antenna, the issue of spatial multiplexing of several substreams onto a number of antennas does not arise.

In this case, the signal model of (1) reduces to

$$\mathbf{x}(k) = \mathbf{h}s(k) + n(k) \qquad (7)$$

where $\mathbf{h} = [h_1 \cdots h_N]^T$ is of dimension $N \times 1$, and both $s(k)$, $n(k)$ are scalars. By evaluating the open-loop capacity (6) for $M = 1$, we obtain (after some algebraic manipulation) the following capacity expression for (flat faded) SIMO systems:

$$C = \log_2 \left(1 + \rho \sum_{n=1}^{N} |h_n|^2 \right) \quad \text{(bps/Hz)} \qquad (8)$$

By contrasting the expression in (8) with the one of (1, 1) systems given in (2), it is clear that, from a capacity perspective, the use of the extra $N - 1$ receiver antennas has resulted in an increase of the link's instantaneous SNR from $\rho|h|^2$ to $\rho \sum_{n=1}^{N} |h_n|^2$. The following observations can be made at this point:

5.1. Capacity Scaling

By rewriting the SNR gain in (8) as

$$\sum_{n=1}^{N} |h_n|^2 = N \times \left(\frac{1}{N} \sum_{n=1}^{N} |h_n|^2 \right)$$

we see that as N grows, the quantity $\frac{1}{N} \sum_{n=1}^{N} |h_n|^2$ converges to an average value. However, the multiplicative factor N keeps growing. In other words, for large N, $C_{1,N}$ grows approximately as

$$C_{1,N} \simeq \log_2(1 + \rho N) \tag{9}$$

that is, *the capacity of a* $(1, N)$ *system increases approximately logarithmically with the number of receiver antennas* N. Notice that in (9), we have assumed that $E|h_n|^2 = 1$ for all $n \in \{1, \ldots, N\}$, which gives $\lim_{N \to \infty} \left(\frac{1}{N} \sum_{n=1}^{N} |h_n|^2 \right) = 1$. The fact that the capacity does not saturate as more antennas are added to the receiver comes from the fact that more power is collected at the receiver for the same power budget at the transmitter.

5.2. Diversity Versus Power Gain

The SNR gain indicated by the multiplicative factor $\sum_{n=1}^{N} |h_n|^2$ is an instantaneous quantity. However, as mentioned above, statistical behavior is important in determining the performance of the system. The statistical behavior is in turn strongly dependent on the degree of correlation between the N receiver antenna elements. The following two extreme cases are typically considered:

- *Uncorrelated Antenna Elements.* This case arises (with good approximation) when the antenna elements are spaced far from each other (on the order of 10λ, where $\lambda = f/c$ is the carrier wavelength). Mathematically, this is expressed by the fact that $E(\mathbf{x}\mathbf{x}^\dagger)$ is a matrix close to diagonal. From a physical point of view, this property indicates that the signals received on different antenna elements fade independently. In this case, the SNR gain represented by the quantity $\sum_{n=1}^{N} |h_n|^2$ represents not only more collected power but also an improvement in the received signal's statistics against fading (it is a combined diversity/power gain).
- *Fully Correlated Antenna Elements.* This case arises when the antennas are closely spaced (on the order of $\lambda/2$). Then, in radiofrequencies, a wavefront impinging on the antenna array causes only phase differences on the signal received from the different elements. In other words, the time interval required by the radiowave to propagate across the antenna array is so small (compared to the inverse of the signal's bandwidth) that the antenna elements appear to be fully correlated.[1] As a result, all the antennas fade simultaneously. The gain $\sum_{n=1}^{N} |h_n|^2$ then is a pure SNR gain — it provides no diversity protection, in the sense that, if one antenna is faded, all other antennas will be faded as well. However, in any given channel realization (even if the channel never changes) the presence of N antennas increases the link's SNR by a factor of N (or $10 \log_{10}(N)$ decibels). This alludes to the classic "3 dB power gain" that is associated in coherent antenna systems with the doubling of antenna elements. Notice that in both extreme cases presented above, the average SNR gain (as well as the average capacity gain) is the same. However, the corresponding CDF's are different.

[1] This is sometimes referred to as the "narrowband assumption" in the antenna array literature [1].

5.3. Receiver Options for Capacity Attainment

Adhering to the assumed nondispersive signal model in (7), the optimal receiver is a linear combiner that operates as follows on the received signal to produce soft outputs for detection:

$$y(k) = \mathbf{h}^\dagger \mathbf{x}(k) \tag{10}$$

It is noteworthy that this simple linear receiver, in the case considered, allows the satisfaction of the following optimality criteria simultaneously:

- Maximum-likelihood detection
- Maximum conditional expectation
- Maximum output signal-to-noise ratio
- Minimum mean-squared error

(This is due partly to the assumed white and Gaussian character of the noise.) More importantly, *this linear receiver allows the attainment of the* $(1, N)$ *channel capacity* in (8). By this phrase we mean that, with the use of progressively stronger — such as Turbo (spatially one-dimensional) — encoding of the original input stream, the capacity in (8) will be attained asymptotically. This property is a direct consequence of the fact that, after the combining in (10) takes place, the $(1, N)$ system has been essentially converted to a (1, 1) system, for which state-of-the art encoding techniques that closely approximate channel capacity exist. In the next section, this property will be contrasted with the capacity attainment capabilities of $(M, 1)$ systems.

5.4. Open-Loop Versus Closed-Loop Operation

Another interesting property of SIMO systems is that their open- and closed-loop capacities coincide. This can be easily verified by comparing Eqs. (5) and (6) for $M = 1$. In this case, $\Phi = \phi$ is scalar. The constraint $tr(\phi) = P_T$ then results in $\phi = P_T$. This results in (6) coinciding with (5) in the case $M = 1$. The result is expected largely since with a single-transmitter antenna, the notion of spatial water filling does not apply. It also corroborates the fact, mentioned above, that $(1, N)$ systems can be represented by equivalent (1, 1) systems, for which there is no difference between open-loop and closed-loop capacities, either.

6. (M, 1) SYSTEMS

The narrowband signal model for systems with a single-receiver antenna is given by

$$x(k) = H\mathbf{s}(k) + n(k) \tag{11}$$

where $H^T = [h_1 \cdots h_M]^T$ and $\mathbf{s}(k) = [s_1(k) \cdots s_M(k)]^T$ are of dimension $M \times 1$ and $x(k)$, $n(k)$ are scalar. As mentioned above, in $(M, 1)$ systems, there are significant differences between the open- and the closed-loop cases. We will hence examine the two cases separately.

6.1. Open-Loop (M, 1) Systems

By evaluating (5) for $N = 1$, we obtain

$$C^{o}_{M,1} = \log_2\left(1 + \frac{\rho}{M}\sum_{m=1}^{M}|h_m|^2\right) \quad \text{(bps/Hz)} \tag{12}$$

By contrasting expression (12) to Eq. (8), we observe the trends described in the following paragraphs.

6.1.1. Diversity Gain. The SNR gain in (12) equals $1/M\left(\sum_{m=1}^{M}|h_m|^2\right)$. Assuming uncorrelated antenna elements, this is a pure *diversity gain*. This is to be contrasted with the joint power/diversity gain achieved in the $(1, N)$ case (see Section 5). This is immediately realized when evaluating the expected value of the SNR gain over many realizations. Assuming that $E|h_m|^2 = 1$ for all $m \in \{1, \ldots, M\}$, we obtain

$$E\left(\frac{1}{M}\sum_{m=1}^{M}|h_m|^2\right) = 1$$

In other words, adding more and more antennas at the transmitter side, without the benefit of more than one antenna at the receiver, does not change the average SNR at the receiver. This stems from the fact that the total transmitted power is kept constant at the transmitter, whereas in the $(1, N)$ case, each extra receiver antenna allows us to collect more signal power against the background noise.

The diversity gain, which is due to the assumption of uncorrelated antenna elements, is again expressed through a change in the CDF of the $(M, 1)$ system for each different M, which results in improved outage capacity (particularly at low outages). Moreover, the highest gain is achieved when going from 1 to 2 transmit antennas. As M keeps growing, it becomes smaller and eventually it saturates. Indeed, in the limit as $M \to \infty$ (and always keeping the total transmit power constant and equal to P_T), the capacity expression in (12) converges to the following expression:

$$C_{\infty,1} = \log_2(1 + \rho) \quad \text{(bps/Hz)} \tag{13}$$

[compare to (9)]. This means that, *in MISO systems, beyond a certain point, there is no benefit in adding more antennas at the transmitter*. Finally, we should note that, if the M antenna elements were fully correlated, there would be no benefit in having $M > 1$, since the received signal's CDF would be the same for all M.

6.1.2. Capacity Attainment. Unlike the simple type of processing required in $(1, N)$ systems to attain (in the sense mentioned above) their capacity in Eq. (8), the attainment of the open-loop capacity (12) in the $(M, 1)$ case seems to be a challenging task. More specifically, attaining the capacity in (12) requires sophisticated space–time coding (STC) techniques [5] at the transmitter. This means that the encoding/spatial multiplexing operations shown in Fig. 1 are nontrivial. So far, only the (2, 1) case seems to admit a straightforward STC technique that allows the attainment of its open-loop capacity [6,7]. This technique, which is briefly described below, relies on a smart (2, 1) space–time multiplexing idea developed by Alamouti [8].[2] In the case $M > 2$, no simple open-loop techniques are known that allow us to attain the capacity in Eq. (12). A (4, 1) technique presented in [9] allows us to get very close to the (4, 1) capacity (achieving on the order of 95% of it or so), and is also briefly discussed below. Despite the diminishing returns of $(M, 1)$ systems for M beyond four antennas, the quest for open-loop capacity attainment is ongoing.

6.1.3. Examples of (M, 1) Space–Time Transmission Schemes

6.1.3.1. (2, 1) Systems: The Alamouti Scheme. An ingenious transmit diversity scheme for the (2, 1) case was introduced by Alamouti [8], and remains to date the most popular scheme for (2, 1) systems. We denote by $\mathbf{S}$ the 2×2 matrix whose (i, j) element is the encoded signal going out of the jth antenna at odd $(i = 1)$ or even $(i = 2)$ time periods (the length of each time period equals the duration of one encoded symbol). In other words, one could think of the vertical dimension of $\mathbf{S}$ as representing "time" and of its horizontal dimension as representing "space." We also denote, as described in Section 2, by $\{b_l(k)\}$, $l = 1, 2$, the encoded version of the lth substream of the original signal $\{\tilde{b}(i)\}$ (see Fig. 1). The Alamouti scheme transmits the following signal every two encoded symbol periods:

$$\mathbf{S}(k) = [\mathbf{s}_1(k)\,\mathbf{s}_2(k)] = \begin{bmatrix} b_1(k) & b_2(k) \\ b_2^*(k) & -b_1^*(k) \end{bmatrix} \tag{14}$$

Having assumed, as noted earlier, the channel to be flat in frequency, the (2, 1) channel is characterized through $H = [\,h_1 \quad h_2\,]$. We group the odd and even samples of the received signal in a 2×1 vector $\mathbf{x}(k)$, which can be then expressed in baseband as

$$\begin{aligned}\mathbf{x}(k) = {} & (h_1(b_1(k)\mathbf{c}_1 + b_2^*(k)\mathbf{c}_2) \\ & + h_2(b_2(k)\mathbf{c}_1 - b_1^*(k)\mathbf{c}_2)) + \mathbf{n}(k)\end{aligned} \tag{15}$$

where $\mathbf{c}_1^T = [1 \mid 0]$, $\mathbf{c}_2^T = [0 \mid 1]$. After subsampling at the receiver and complex-conjugating the second output, we obtain

$$\begin{aligned} d_1(k) &= \mathbf{c}_1^T\mathbf{x}(k) = (h_1 b_1(k) + h_2 b_2(k)) + v_1(k) \\ d_2(k) &= (\mathbf{c}_2^T\mathbf{x}(k))^* = (-h_2^* b_1(k) + h_1^* b_2(k)) + v_2^*(k) \end{aligned} \tag{16}$$

[2] An extension of this technique to CDMA systems, called space–time spreading (STS), was presented in [11] and has been introduced in third-generation wireless standards for the (2, 1) case.

where $v_m(k) = \mathbf{c}_m^T \mathbf{n}(k)$, $m = 1, 2$. Equation (16) can be equivalently written as

$$\begin{aligned}\mathbf{d}(k) &= \begin{bmatrix} h_1 & h_2 \\ -h_2^* & h_1^* \end{bmatrix} \begin{bmatrix} b_1(k) \\ b_2(k) \end{bmatrix} + \mathbf{v}(k) \\ &= \mathbf{H}\mathbf{b}(k) + \mathbf{v}(k)\end{aligned} \tag{17}$$

where $\mathbf{v}^T(k) = [\, v_1(k) \quad v_2^*(k) \,]$ and $\mathbf{H}$ is a unitary matrix (up to a complex scalar). After match filtering to $\mathbf{H}$, we obtain

$$\begin{aligned}\mathbf{d}'(k) = \mathbf{H}^\dagger \mathbf{d}(k) = &\begin{bmatrix} |h_1|^2 + |h_2|^2 & 0 \\ 0 & |h_1|^2 + |h_2|^2 \end{bmatrix} \\ &\times \mathbf{b}(k) + \mathbf{v}'(k)\end{aligned} \tag{18}$$

where $\mathbf{v}'(k)$ remains spatially white. Because of the diagonal character of the mixing matrix in (18), the 2×1 space–time system has been now reduced to an equivalent set of two 1×1 systems! We call this the *decoupled property* of a space–time code. Moreover, each of these two equivalent single-dimensional systems has the same capacity. We call this the property of *balance* of a space–time code.

The total constrained capacity of this system equals the sum of the capacities of the two SISO systems (each SISO system operates at half the original information rate):

$$C_{2,1}^A = \log_2 \left(1 + \frac{\rho}{2}(|h_1|^2 + |h_2|^2)\right) \tag{19}$$

By contrasting (19) to (12) with $M = 2$, we see that

$$C_{2,1}^A = C_{2,1}^o \tag{20}$$

Hence, *the Alamouti scheme allows the attainment of the* $(2, 1)$ *open-loop capacity*. Moreover, this is possible with the use of conventional (spatially single-dimensional, such as Turbo) encoding. The Alamouti scheme remains, to the best of our knowledge, the only decoupled and balanced space–time code that allows the attainment of the system's full open-loop capacity.

6.1.3.2. (4, 1) Systems: A More Recently Proposed Scheme. As mentioned above, the attainment of the open-loop capacity in the general $(M, 1)$ case is a challenging problem. A scheme for the (4, 1) case that approaches the capacity closely was proposed in 2001 [9]. According to this scheme, the original information sequence $\tilde{b}(i)$ is first demultiplexed into four encoded substreams $b_m(k)$ $(m = 1, \ldots, 4)$. The four-dimensional transmitted signal is then organized in blocks of $L = 4$ (encoded) symbol periods and is represented by a 4×4 matrix $\mathbf{S}$, which is arranged as follows:

$$\mathbf{S} = \begin{bmatrix} b_1 & b_2 & b_3 & b_4 \\ b_2^* & -b_1^* & b_4^* & -b_3^* \\ b_3 & -b_4 & -b_1 & b_2 \\ b_4^* & b_3^* & -b_2^* & -b_1^* \end{bmatrix} \tag{21}$$

where the time index k has been dropped for convenience. Similar to the (2, 1) case, the mth column of $\mathbf{S}$ in (21) represents a block of 4 symbols that are transmitted from the mth transmit antenna. The maximum capacity attainable by this transmission technique was computed in [9] and it is given by the following expression:

$$C_{4,1}^{\text{proposed,max}} = \frac{1}{2} \log_2 \det \left(\mathbf{I}_2 + \frac{\rho}{4} \Delta_1\right) \tag{22}$$

where

$$\Delta_1 = \begin{bmatrix} \gamma & \alpha \\ -\alpha & \gamma \end{bmatrix} \tag{23}$$

and

$$\begin{aligned}\gamma &= \mathbf{h}^H \mathbf{h} = \sum_{m=1}^{4} |h_m|^2 \\ \alpha &= 2j\text{Im}(h_1^* h_3 + h_4^* h_2)\end{aligned} \tag{24}$$

(where Im denotes the imaginary part of a complex scalar). Some quantitative results regarding the capacities of these techniques will be given in Section 7.

6.2. Closed-Loop $(M, 1)$ Systems

Unlike the open-loop case, the closed-loop capacity expression for $(M, 1)$ systems is given by the following expression:

$$C_{M,1}^c = \log_2 \left(1 + \rho \sum_{m=1}^{M} |h_m|^2\right) \quad \text{(bps/Hz)} \tag{25}$$

Notice that the SNR gain in Eq. (25) is similar to the $(1, N)$ case [see Eq. (8)]. The use of extra transmitter antennas now adds to the receiver power (and hence capacity keeps growing with M). For large M, the closed-loop capacity of $(M, 1)$ systems scales as follows:

$$C_{M,1} \simeq \log_2(1 + \rho M) \quad \text{(bps/Hz)} \tag{26}$$

[compare to Eq. (9)]. When the channel H is fully known at the transmitter, the closed-loop capacity in (25) is easily attainable through spatial waterfilling at the transmitter, as will be described below.

In conclusion, we observe that $(N, 1)$ and $(1, N)$ systems are not in general symmetric. They are, however, symmetric, when the $(N, 1)$ channel is perfectly known at the transmitter, allowing it to perform spatial maximal ratio combining (MRC) before transmission (spatial pre-equalization), as will be shown below.

6.2.1. Example Schemes

6.2.1.1. Transmit MRC. The optimal transmission approach in the $(M, 1)$ case, when the channel is flat and fully known at the transmitter, is to send the following signal out of the M antennas:

$$\mathbf{s}(k) = \left(\frac{1}{\sqrt{\sum_{m=1}^{M} |h_m|^2}}\right) \begin{bmatrix} h_1^* \\ \vdots \\ h_M^* \end{bmatrix} b(k) \tag{27}$$

(see [11]) which amounts to performing MRC at the transmitter. The operation is mathematically equivalent to the receive MRC performed at the receiver in $(1, N)$ systems, as described in (10). Notice that the same information sequence is sent simultaneously out of all the antennas, however it is multiplied by a different complex scalar on each antenna. As mentioned above, this

simple transmission scheme (provided that the channel is perfectly known at the transmitter), achieves the $(M, 1)$ closed-loop capacity for any M.

6.2.1.2. Switch Transmit Diversity (STD). In some cases, good knowledge of the channel coefficients at the base station is not feasible. This may be due, for example, to the excessive amount of feedback required in order to send back reliably the channel information from the terminal to the base station, the highly time-varying nature of the channel (high Doppler), errors in channel estimation, and errors in the feedback channel. In such cases, there is interest in exploring alternative methods that make use of *partial* channel state information at the transmitter. One very good and simple candidate is so-called selection transmit diversity (STD). This technique transmits, at each symbol period, only from one antenna, at full power. The symbol is transmitted from the antenna that experiences the highest SNR during that symbol period. Mathematically, STD transmission can be described as follows:

$$\mathbf{s}(k) = \delta_m(k) b(k) \tag{28}$$

where $\delta_m(k)$ is an $M \times 1$ vector, whose single nonzero entry is at position m, where $m \in \{1, \ldots, M\}$ is such that $|h_m|$ takes the highest value in that set during the kth symbol period.

From a practical point of view, this technique is appealing because it requires only the knowledge of which is the strongest antenna element ($\log_2(M)$ bits of information per symbol). From a capacity point of view, the STD scheme is capable of achieving the following capacity:

$$C_{M,1}^{\text{STD}} = \log_2(1 + \rho \max_{m \in \{1,\ldots,M\}} |h_m|^2) \quad \text{(bps/Hz)} \tag{29}$$

As it turns out, the capacity in (29) typically lies roughly midway between the $(M, 1)$ open-loop capacity (12) and the $(M, 1)$ closed-loop capacity (25).

6.2.2. Numerical Examples. Figure 3 shows the 10% outage capacities for some open- and closed-loop $(M, 1)$ cases. As predicted by expressions (12) and (25), the difference between the (4, 1) open- and closed-loop capacities is about 6 dB. Moreover, notice that the (4, 1) open-loop capacity is only about 2 dB better than the (2, 1) open-loop capacity (diminishing returns), whereas the (4, 1) STD capacity lies midway between the open- and closed-loop capacities.

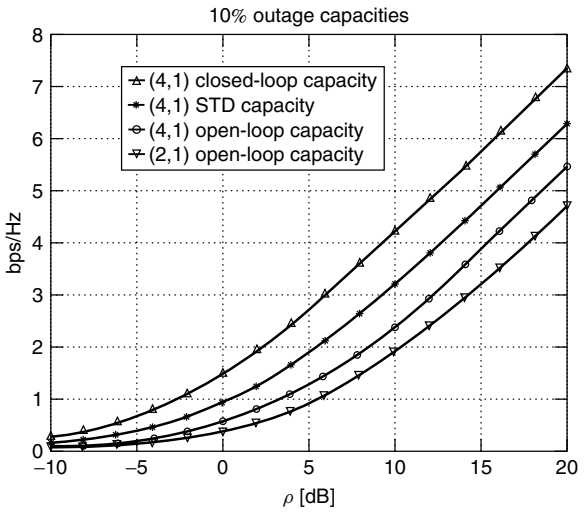


Figure 3. Open- and closed-loop $(M, 1)$ capacities.

7. (M, N) SYSTEMS

In this section we will present some practical existing techniques that attempt to approach the capacity of open-loop systems in the general (M, N) case.

7.1. Combined *Transmit* + *Receive* Diversity Systems

Given a certain $(M, 1)$ system, one straightforward way to design an (M, N) system is to simply

- Transmit as in the $(M, 1)$ system
- Receive on each antenna as in the $(M, 1)$ system
- Combine optimally the N receiver antenna outputs

The capacity quantification of these transmit/receive diversity systems is straightforward. The $M \times N$ (assumed flat) channel is represented through the $N \times M$ channel matrix:

$$\mathbf{H} = \begin{bmatrix} h_{11} & \cdots & h_{1M} \\ \vdots & \ddots & \vdots \\ h_{N1} & \cdots & h_{NM} \end{bmatrix} = \begin{bmatrix} \mathbf{h}_1 & \cdots & \mathbf{h}_N \end{bmatrix}$$

We first compute an upper bound for the capacity of such an (M, N) transmit/receive diversity system. With optimal ratio combining, and assuming that each $(M, 1)$ system takes no interference hit, the input/output relationship takes the form

$$\mathbf{d}(k) = \left(\sum_{m=1}^{M} \sum_{n=1}^{N} |h_{nm}|^2 \right) \mathbf{b}(k) + \mathbf{n}(k) \tag{30}$$

where $\mathbf{d}(k)$, $\mathbf{b}(k)$, and $\mathbf{n}(k)$ are all of dimension $M \times 1$. The corresponding capacity is given by

$$C_{M,N}^{\text{trd,max}} = \log_2 \left(1 + \frac{\rho}{M} \sum_{m=1}^{M} \sum_{n=1}^{N} |h_{nm}|^2 \right) \tag{31}$$

It is clear that, when the attainable capacity of the corresponding $(M, 1)$ schemes is away from the $(M, 1)$ log-det capacity, the upper bound in (31) will not be attained either. Notice further that the expression in (31) is strictly smaller than the (M, N) log-det capacity in (5) for $N > 1$.

7.1.1. Example Schemes. To give some examples, the capacity of a $(2, N)$ system that uses the Alamouti (2, 1) scheme is

$$C_{2,N}^{A} = \log_2 \left(1 + \frac{\rho}{2} \sum_{n=1}^{N} (|h_{n,1}|^2 + |h_{n,2}|^2) \right) \tag{32}$$

thus, as expected, the upper bound in (31) is "attained" by the Alamouti scheme in the $(2, N)$ case. However, this still falls short of the $(2, N)$ log-det capacity (5).

It is also straightforward to compute the maximum attainable capacity of a $(4, N)$ system that uses the (4, 1) scheme of [9], which is given by

$$C_{4,N}^{\text{proposed,max}} = \frac{1}{2}\log_2 \det\left(\mathbf{I}_2 + \frac{\rho}{4}\mathbf{\Gamma}_{2N}\mathbf{\Gamma}_{2N}^{\dagger}\right) \tag{33}$$

where $\mathbf{\Gamma}_{2N} = [\Gamma_1^T \cdots \Gamma_N^T]^T$, with Γ_n defined from

$$\Delta_n = \Gamma_n \Gamma_n^{\dagger}$$

and where Δ_n is defined similarly to Δ_1 in (23) for the nth (as opposed to the first) receiver antenna.

7.1.2. Numerical Examples. Figures 4 and 5 show 10% outage capacities that were numerically evaluated for some schemes based on the capacity expressions that were presented above. All expressions were run over an ensemble of 10^4 (M, N) random Rayleigh-faded channel matrices (each entry of the matrix is chosen independently from any other entry from a complex i.i.d. Gaussian distribution of unit variance). The 10% outage value was then selected from the corresponding point of the CDF.

Figure 4 shows the 10% outage capacities for several $(M, 1)$ cases, as well as for the (2, 2) case. In the (2, 1) case, the plotted capacity corresponds both to the Alamouti scheme and to the maximum open-loop capacity, as indicated by Eq. (20). For the other $(M, 1)$ cases, we plot the capacity upper bounds corresponding to Eq. (12), and we use Eq. (13) for the asymptotic $(\infty, 1)$ case. We also use Eq. (31) with $N = 2$ for the capacity of a (2, 2) combined Alamouti/receive diversity scheme, and the log-det expression (5) for the (2, 2) maximum open-loop capacity. We observe that, at 10 dB, *the* $(2, 1)$ *system almost doubles the capacity of the* $(1, 1)$ *system*! However, as noted earlier, further increasing the number of transmit antennas in the $(M, 1)$ case offers diminishing returns. It is also worth noting that the (2, 2) combined transmit/receiver diversity scheme is capable of attaining

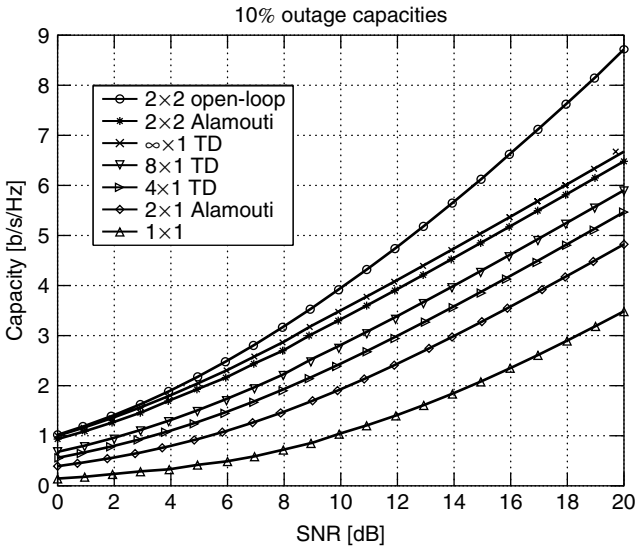


Figure 4. Outage capacities and bounds of $(M, 1)$ and $(M, 2)$ schemes.

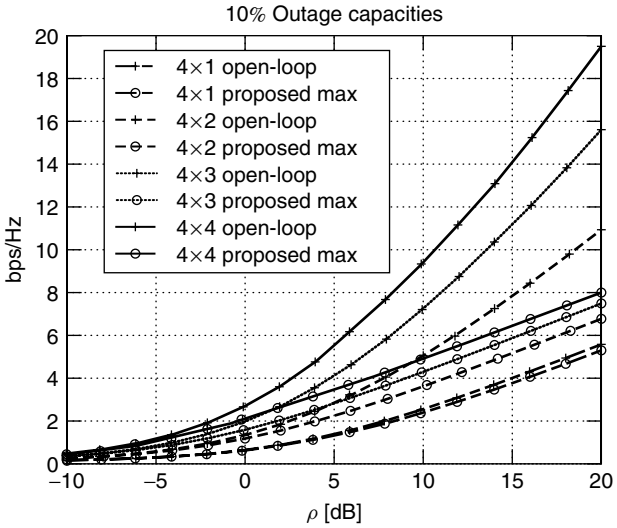


Figure 5. Outage capacities of the (4, 1) scheme described in Ref. 9, when used with up to four receiver antennas.

a quite significant fraction (particularly at low SNRs) of the maximum (2, 2) open-loop capacity. Finally, it is also interesting to note that a (2, 2) system achieves about the same capacity as an open-loop $(\infty, 1)$ system, which conveys again the message of the high value of adding extra antennas at the receiver.

In Fig. 5, we show the capacities of some combined transmit/receive diversity schemes for different $(4, N)$ cases. The circles represent the combined $(4, N)$ systems corresponding to the (4, 1) scheme of [9], in conjunction with optimal receiver diversity. When read from the bottom up, these four curves correspond to $N = 1, 2, 3, 4$, respectively. Similarly, the crosses represent the corresponding open-loop $(4, N)$ capacities. Notice that the proposed (4, 1) scheme is very close to the open-loop capacity; however, the gap gets increasingly larger as N grows from 1 to 4. In the (4, 2) case though, the scheme still performs well, particularly at low SNRs.

7.2. V-BLAST

A quite simple, from the transmitter's point of view, space–time transmission scheme was proposed in [10], and it is widely referred to as "V-BLAST," which stands for *Vertical Bell labs LAyered Space–Time*. In this architecture, $\{\tilde{b}(i)\}$ is first demultiplexed into M substreams, which are then encoded independently and mapped each on a different antenna:

$$s_m(k) = b_m(k), \qquad m = 1, \ldots, M$$

In other words, the original bit stream is converted into a vertical vector of encoded substreams (whence the term "vertical" BLAST), which are then streamed to the antennas through a 1–1 mapping. In [10], it was proposed to process the received signal with the use of a successive interference canceller. After determining the order into which the M substreams will be detected, the V-BLAST

receiver operates according to the following generic three-stage scheme, which is performed in a successive fashion for each substream:

1. Project away from the remaining interfering substreams.
2. Detect (after decoding, deinterleaving, and slicing) the substream.
3. Cancel the effect of the detected substream from subsequent substreams.

Mathematically, these operations can be described as follows for the k_mth substream:

$$\begin{aligned} z_{k_m}(k) &= W_{k_m}^{\dagger}\mathbf{x}^m(k) \\ \hat{z}_{k_m}(i) &= \mathrm{dec}(z_{k_m}(k)) \\ \mathbf{x}^{m+1}(k) &= \mathbf{x}^m(k) - \mathrm{enc}(\hat{z}_{k_m}(i))\mathbf{h}_{k_m} \end{aligned} \tag{34}$$

where $\mathbf{x}^1(k) = \mathbf{x}(k)$, $\{k_1, \ldots, k_M\}$ is a reordered version of the set $\{1, \ldots, M\}$ that determines the order in which the substreams will be detected, $\mathrm{dec}(\cdot)$ represents the decoding+detection operation, and $\mathrm{enc}(\cdot)$ represents the encoding operation. Finally, W_{k_m} represents the $N \times 1$ vector that operates on $\mathbf{x}^m(k)$ in order to project away from substreams $\{k_{m+1}, \ldots, k_M\}$. The operations in Eq. (34) are performed successively for $m = 1, \ldots, M$, after the ordering $\{k_1, \ldots, k_M\}$ has been determined.

We now discern between the following two cases for this linear operation, since they affect significantly the constrained capacity of the system:

7.2.1. Zero-Forcing Projection. In this case, at the mth stage, $W_{k_m}^{\dagger}$ nulls perfectly the interference from all the remaining (undetected) substreams. These are the substreams with indices $\{k_{m+1}, \ldots, k_M\}$. This nulling is represented mathematically as

$$W_{\mathrm{zf},k_m}^{\dagger}\mathbf{H} = [0 \cdots 0\ 1\ 0 \cdots 0] = \delta_{k_m}^T \tag{35}$$

where the unique non-zero element of the $1 \times M$ vector δ_{k_m} is in its k_mth position. As a result, the end-to-end model for the k_mth output is

$$d_{k_m}(k) = b_{k_m}(k) + W_{\mathrm{zf},k_m}^{\dagger}\mathbf{n}(k), m = 1, \ldots, M \tag{36}$$

where $\mathbf{n}(k) = [n_1(k) \cdots n_N(k)]^T$ is the receiver noise. Defining $\mathbf{d}(k) = [d_{k_1}(k) \cdots d_{k_M}(k)]^T$ and $\mathbf{b}(k) = [b_{k_1}(k) \cdots b_{k_M}(k)]^T$, (36) can be written in matrix form as

$$\mathbf{d}(k) = \mathbf{b}(k) + \mathbf{W}_{\mathrm{zf}}^{\dagger}\mathbf{n}(k) \tag{37}$$

where $\mathbf{W}_{\mathrm{zf}} = [W_{\mathrm{zf},k_1} \cdots W_{\mathrm{zf},k_M}]$. From (37), we observe that the ZF version of the V-BLAST superstructure is

- *Decomposable* — its capacity can be evaluated by computing separately each capacity of its M substreams.
- *Unbalanced* — all the substreams have different capacities, reflecting the fact that each "sees" a different SNR (this is also reflected in the fact that different columns of $\mathbf{W}_{\mathrm{zf}}$ have in general different square norms).

These attributes may be contrasted with the Alamouti (or STS) architecture mentioned in Section 6.1, which is both decomposable and balanced (each of its two substreams is capable of carrying exactly the same information rate). Regarding the capacity of the end-to-end system, it is important to emphasize that we have assumed that each substream is independently encoded, and that the transmitter has no way of knowing which is the maximum attainable rate for each antenna. As a result, it can at best transmit from all antennas the same rate. Hence, the capacity will equal M times the smallest of the M decomposed channel capacities:

$$C_{MN}^{\mathrm{VB-ZF}} = M \times \min_{m \in \{1,\ldots,M\}} \{\log_2(1 + \rho_{\mathrm{zf},k_m})\} \tag{38}$$

where ρ_{k_m} is the output SNR of the k_mth substream:

$$\rho_{\mathrm{zf},k_m} = \frac{\rho}{M|W_{\mathrm{zf},k_M}|^2} \tag{39}$$

It should finally be noted that the capacity in Eq. (38) can be optimized by choosing an optimal ordering for the set $\{k_1, \ldots, k_M\}$ [10].

7.2.2. MMSE Projection. In this case, at the mth stage, an optimal compromise between linear interference mitigation of the undetected substreams and noise amplification is sought. This is achieved through the following minimum mean squared error (MMSE) criterion:

$$\min_{W_{k_m}} E\|d_{k_m} - W_{k_m}^{\dagger}\mathbf{H}_{k_m}\|^2 \tag{40}$$

where $\mathbf{H}_{k_m}$ is derived from $\mathbf{H}$ by deleting its columns corresponding to indices $\{k_1, \ldots, k_{m-1}\}$. This gives for W_{k_m}:

$$W_{\mathrm{mmse},k_m}^{\dagger} = \left(\mathbf{H}_{k_m}\mathbf{H}_{k_m}^{\dagger} + \frac{M}{\rho}\mathbf{I}_N\right)^{-1}\mathbf{h}_{k_m} \tag{41}$$

where $\mathbf{h}_{k_m}$ is the k_mth column of $\mathbf{H}$. This end-to-end system has again been fully decomposed into four (1, 1) systems, which are, in general, not balanced (i.e., they do not have the same SINRs). Its capacity is computed again through the minimum of the four 1×1 capacities, and is given by a formula similar to (38):

$$C_{MN}^{\mathrm{VB-MMSE}} = M \times \min_{m \in \{1,\ldots,M\}} \{\log_2(1 + \rho_{\mathrm{mmse},k_m})\} \tag{42}$$

where now

$$\rho_{\mathrm{mmse},k_m} = \frac{\|W_{\mathrm{mmse},k_m}^{\dagger}\mathbf{H}_{k_m}\|^2}{M\|W_{\mathrm{mmse},k_m}\|^2/\rho + \sum_{l \neq k_m}\|W_{\mathrm{mmse},l}\|^2} \tag{43}$$

Again, the capacity in Eq. (42) can be maximized through optimal ordering.

7.2.3. Closed-Loop V-BLAST Operation. The fact that in both (38) and (42) the system capacity is a multiple of the

weakest rate is a direct result of the absence of knowledge of these rates at the transmitter. Had we assumed that, indeed, all M rates were known at the transmitter, the expressions (38) and (42) would use the sum of capacities as opposed to M times the minimum of capacities. A quite astonishing result that was reported in [12] (based on previous work in [13]) is that the sum of MMSE capacities in (42) equals the open-loop capacity of the MIMO channel in (5)! In other words

$$\sum_{m\in\{1,\ldots,M\}} \{\log_2(1+\rho_{\mathrm{mmse},k_m})\} = \log_2\left\{\det\left(I_N + \frac{\rho}{M}\mathbf{H}\mathbf{H}^\dagger\right)\right\} \tag{44}$$

This means, that *if the maximum attainable rates are known at the V-BLAST transmitter, then the system can attain the open-loop capacity with the use of linear MMSE* (+*subtractions*) *processing*. Moreover, the result holds irrespective of the ordering of the substreams. This is a quite nice tradeoff: a partially closed-loop technique (the receiver only feeds back to the transmitter a set of rates) allows to attain the system's open-loop capacity!

7.2.4. Numerical Results. In Fig. 6, we show a capacity CDF, at 10 dB SNR, of the ZF and MMSE V-BLAST architectures described above for the (4, 4) case. Notice that, at this SNR, the MMSE architecture is capable of attaining about 70% of the total open-loop capacity at 10% outage. However, the ZF architecture performs poorly, and it is even outperformed by a (1, 4) maximal ratio combining system at outages higher than 20%! The situation is more severe for lower SNRs such as 0 dB, as shown in Fig. 7. Now the V-BLAST MMSE architecture attains only about 50% of the (4, 4) open-loop capacity, whereas the ZF architecture is outperformed by the (1, 4) system across the board.

7.3. Other (M, N) Schemes

Similar to the (M, 1) case, several other schemes have been proposed in the literature for the general (M, N) case. For example, the use of a block space–time multiplexing whose

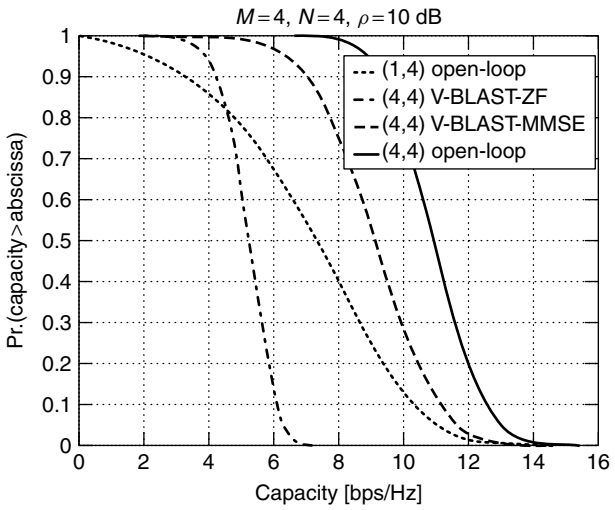


Figure 6. Outage capacity distribution of a V-BLAST MMSE architecture at 10 dB SNR.

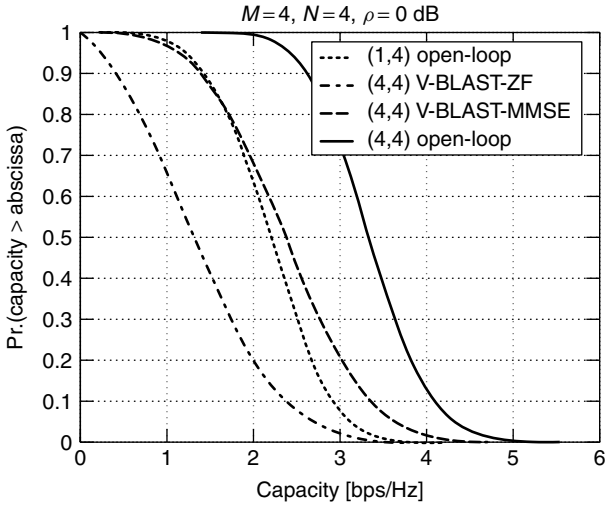


Figure 7. Outage capacity distribution of a V-BLAST MMSE architecture at 0 dB SNR.

mixing coefficients are optimized numerically according to a maximum *average capacity* criterion has been suggested in [14]. Another approach in [15] uses Turbo codes in the following way. The original substream is first demultiplexed into M substreams, which are separately encoded, each with a block code. Then, the M encoded outputs are space–time-interleaved in a random fashion, mapped onto constellation symbols, and sent out of the M antennas. At the receiver, the M substreams are separated through an iterative interference canceler, which uses MMSE for the linear (soft) part, and subtracts decisions made after (joint) deinterleaving and (separate) decoding of each interfering substream in the cancellation part.

These approaches have demonstrated encouraging performance in terms of bit/frame error rate at the receiver. However, their inherent capacity penalties are still unknown, mainly because of their apparent lack of structure and other properties such as the ones discussed above. The quantification of the capacity penalties of these and other emerging space–time transmission techniques remains an interesting open question.

8. CONCLUSIONS

In this article, we have taken a capacity view of multiple antenna wireless systems. We have outlined the fundamental channel capacity formulas that govern the spectral efficiencies of such multiple-input/multiple-output (MIMO) systems. We then described how these expressions reduce in a number of special cases. This has allowed us to draw interesting interpretations regarding the potential of different existing techniques to approximate the capacities promised by the formulas. Moreover, we believe that the capacity view has shed some new light on understanding the value of more conventional antenna combining techniques. Besides helping assessing the value of existing techniques, we believe that these results can be used to identify directions for future research in the field of MIMO systems.

Acknowledgments
The author would like to thank Dr. G. J. Foschini for many helpful and exciting discussions on the topic of MIMO systems, as well as his many colleagues from Bell Labs' Wireless Research Lab for numerous fruitful interactions on the topic.

BIOGRAPHY

Constantinos Papadias was born in Athens, Greece, in 1969. He received the diploma of electrical engineering from the National Technical University of Athens (NTUA), Greece, in 1991 and a Ph.D. degree in signal processing (highest honors) from the Ecole Nationale Supérieure des Télécommunications (ENST), Paris, France, in 1995. From 1992 to 1995 he was a teaching and research assistant at the Mobile Communications Department, Eurécom, France. In 1995, he joined the Information Systems Laboratory, Stanford University, California, as a Postdoctoral researcher, working in the Smart Antennas Research Group. In November 1997 he joined the Wireless Research Laboratory of Bell Labs, Lucent Technologies, Holmdel, New Jersey, as a member of the technical staff. He is now a technical manager in Bell Laboratories Global Wireless Systems Research Department. His current research interests lie in the areas of multiple antenna systems (e.g., MIMO transceiver design and space-time coding), interference mitigation techniques, reconfigurable wireless networks, as well as financial evaluation of wireless technologies. He has authored several papers and patents on these topics. Dr. Papadias is a member of IEEE and a member of the Technical Chamber of Greece.

BIBLIOGRAPHY

1. A. Paulraj and C. Papadias, Space-time processing for wireless communications, *IEEE Signal Process. Mag.* **14**(6): 49–83 (Nov. 1997).
2. G. J. Foschini, Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas, *Bell Labs Tech. J.* **1**(2): 41–59 (1996).
3. E. Telatar, *Capacity of Multi-antenna Gaussian Channels*, AT & T Bell Laboratories Technical Memorandum, June 1995.
4. G. Foschini and M. Gans, On limits of wireless communications in a fading environment when using multiple antennas, *Wireless Pers. Commun.* **6**(6): 315–335 (1998).
5. V. Tarokh, N. Seshadri, and A. R. Calderbank, Space-time codes for high data rate wireless communication: Performance criterion and code construction, *IEEE Trans. Inform. Theory* **44**(2): 744–765 (March 1998).
6. C. Papadias, On the spectral efficiency of space-time spreading schemes for multiple antenna CDMA systems, *33rd Asilomar Conf. Signals, Systems, and Computers*, Pacific Grove, CA, Oct. 24–27, 1999, pp. 639–643.
7. S. Sandhu and A. Paulraj, Space-time block codes: A capacity perspective, *IEEE Commun. Lett.* **4**(12): 384–386 (Dec. 2000).
8. S. Alamouti, A simple transmitter diversity scheme for wireless communications, *IEEE J. Select. Areas Commun.* **16**: 1451–1458 (Oct. 1998).
9. C. Papadias and G. J. Foschini, A space-time coding approach for systems employing four transmit antennas, *Int. Conf. Acoustics, Speech, and Signal Processing* (*ICASSP 2001*), Salt Lake City, UT, May 7–11, 2001.
10. G. J. Foschini, G. D. Golden, R. A. Valenzuela, and P. W. Wolniansky, Simplified processing for wireless communication at high spectral efficiency, *IEEE J. Select. Areas Commun.* **17**(11): 1841–1852 (Nov. 1999).
11. B. Hochwald, L. Marzetta, and C. Papadias, A transmitter diversity scheme for wideband CDMA systems based on space-time spreading, *IEEE J. Select. Areas Commun.* **19**(1): 48–60 (Jan. 2001).
12. S. T. Chung and A. Lozano, and H. C. Huang, Approaching eigenmode BLAST channel capacity using V-BLAST with rate and power feedback, *Vehicular Technology Conf.* (*VTC*) *Fall 2001*, Atlantic City, NJ, Oct. 2001.
13. M. K. Varanassi and T. Guess, Optimum decision-feedback multiuser equalization with successive decoding achieves the total capacity of the Gaussian multiple-access channel, *1998 Asilomar Conf. Signals, Systems, and Computers*, 1998, pp. 1405–1409.
14. B. Hassibi and B. Hochwald, High-rate linear space-time codes, *Int. Conf. Acoustics, Speech, and Signal Processing* (*ICASSP 2001*), Salt Lake City, UT, May 7–11, 2001.
15. M. Sellathurai and S. Haykin, Joint beamformer estimation and co-antenna interference cancelation for Turbo-BLAST, *Int. Conf. Acoustics, Speech, and Signal Processing* (*ICASSP 2001*), Salt Lake City, UT, May 7–11, 2001.

MULTIPROTOCOL LABEL SWITCHING (MPLS)

Pim Van Heuven
Steven Van Den Berghe
Filip De Turck
Piet Demeester
Ghent University
Ghent, Belgium

1. INTRODUCTION

The Internet Protocol (IP) is a connectionless networking layer protocol to route IP packets over a network of IP routers. Every router in the IP network examines the IP packet header and independently determines the next hop based on its internal routing table. A routing table contains information about the next hop and the outgoing interface for the destination address of each IP address. Multiprotocol label switching (MPLS) allows the setup of label-switched paths (LSPs) between IP routers to avoid the IP forwarding in the intermediate routers. IP packets are tagged with labels. The initial goal of label-based switching was to increase the throughput of IP packet forwarding. Label-based switching methods allow routers to make forwarding decisions based on the contents of a simple label, rather than by performing a complex route lookup according to the destination IP address. This initial justification for MPLS is no longer perceived as the main benefit, since nowadays routers are able to perform route lookups at sufficiently high speeds to support most interface types. However, MPLS brings many other benefits to IP-based networks, including

(1) traffic engineering, that is, the optimization of traffic handling in networks; (2) virtual private networks (VPNs), networks that offer private communication over the public Internet using secure links; and (3) the elimination of multiple protocol layers.

MPLS paths are constructed by installing label state in the subsequent routers of the path. Labels are fixed-length entities that have only a local meaning. Labels are installed with a label distribution protocol. MPLS forwarding of the IP packets is based on these labels. The IP and MPLS forwarding principles will be detailed first, followed by a description of the MPLS label distribution process.

1.1. Forwarding in IP and MPLS

In regular non-MPLS IP networks, packets are forwarded in a hop-by-hop manner. This means that the forwarding decision of a packet traversing the network is based on the lookup of the destination in the local routing table [also called *routing information base* (RIB)]. Figure 1a illustrates IP for a network consisting of four routers: nodes A, B, C and D. A simplified IP routing table of router B is shown. It consists of entries that map the destination network addresses of the IP packets to the IP addresses of the next hop and the router interface, which is connected to the next hop. When forwarding a packet, a router inspects the destination address of the packet (found in the IP header), searches through his local router table via a longest prefix match, and forwards it to the next hop on the outgoing interface.

The destination addresses in this table are aggregated in order to reduce the number of entries in this table. These entries are aggregated by indicating the length of the significant part of the destination addresses (from 0 to 32 bits). If n is the length of address a, then only the first n (most significant) bits of a are considered. The resulting partial address is called a *prefix* and is noted as a/n (e.g., 10.15.16.0/24). This aggregation of addresses has the drawback that searching through the table needs to be done with a *longest-prefix* match. A longest-prefix match is more complex than an exact match because the result of the search must be the entry with the longest prefix that matches the address [1].

An important characteristic of IP forwarding is that packets arriving at a router with the same destination prefix are forwarded equivalently over the network. A class of packets that can be forwarded equivalently is a forwarding equivalence class (FEC). Because of the destination-based forwarding of IP, FECs are usually associated with IP prefixes. The forwarding in an IP router can be restated as the partitioning of packets in FECs and assigning a next hop to the FECs. It is important to note that determination of the FEC needs to be done in every hop for every packet.

On the other hand, MPLS forwarding relies on labels, instead of prefixes to route packets through the network [2,3]. Labels are fixed-length entities that have only a local meaning. Because a label has only a local meaning, labels can be different at every hop and therefore must be adapted before forwarding the packet; this process is called *label switching*. The labels are distributed over the MPLS domain by means of a label distribution protocol. MPLS routers are called *label-switching routers* (LSR) (Fig. 1c) because they operate on labels rather than on IP prefixes when forwarding packets. The concatenation of these installed labels in the different LSRs is called a *label-switched patch* (LSP). An LSP is set up between the ingress LSR and the egress LSR; these edge LSRs are also called *label edge routers* (LERs). Packets belonging to a certain FEC are then mapped on an LSP. Determining the FEC of a packet is necessary only in the ingress of the LSP. The segregation of packets in FECs needs be done only once, in the ingress router, and this segregation can also be based on more than the destination prefix of the packet. For example, it is possible to take both the source and the destination into account. LSPs are unidirectional paths because they are based on FEC-to-label bindings.

In the case of label-switched routers, every router contains two tables: an incoming label map (ILM) table that contains all the incoming labels the router has allocated and a table that contains all the necessary information to forward a packet over an LSP (Fig. 1b). The latter table is populated with next-hop label-forwarding entries (NHLFE). There is a mapping between the ILM and an NHLFE mapping the incoming labels to an output label, the outgoing interfaces, and the next hop. The router inspects the incoming label and consults the ILM table to find the right NHLFE. This NHLFE contains the outgoing label, the next hop, and the outgoing interface. Before the packet is sent to the next hop, the label is switched to the outgoing label value.

1.2. LSP Setup

Two distinct cases can be distinguished: (1) hop-by-hop routed LSP setup and (2) explicit routed LSP setup. These two cases will be described in the following sections.

1.2.1. Hop-by-Hop Routed LSP Setup. To distribute labels over the network and consequently set up an LSP, a *label distribution protocol* is used. Path setup typically consists of two steps: (1) a request is sent to the egress of the LSP and (2) the response propagates back to the ingress. The first step is denoted by the generic term "label request," whereas the second step is denoted by the term "label mapping." Figure 2a illustrates the label distribution process. When LER A wants to set up an LSP to network *netD*, it will send a label request to its next hop toward netD (step a, Fig. 2). The intermediate nodes from the ingress towards the egress (like LSR B) will install state about the request and will forward the request toward netD according to their routing information bases (step b). When the request reaches the destination of the LSP, the egress node will allocate a label for this LSP and will store this information in the incoming label map (ILM). The LSR will then send a label mapping back to the previous hop. The "label mapping" message contains the label previously allocated by the LSR (step d). LER B will then receive the label mapping from node D. The label contained in the label mapping will be used to make a next-hop label-forwarding entry (NHLFE). Router B will then, in turn, allocate a label and store this label in its ILM. The

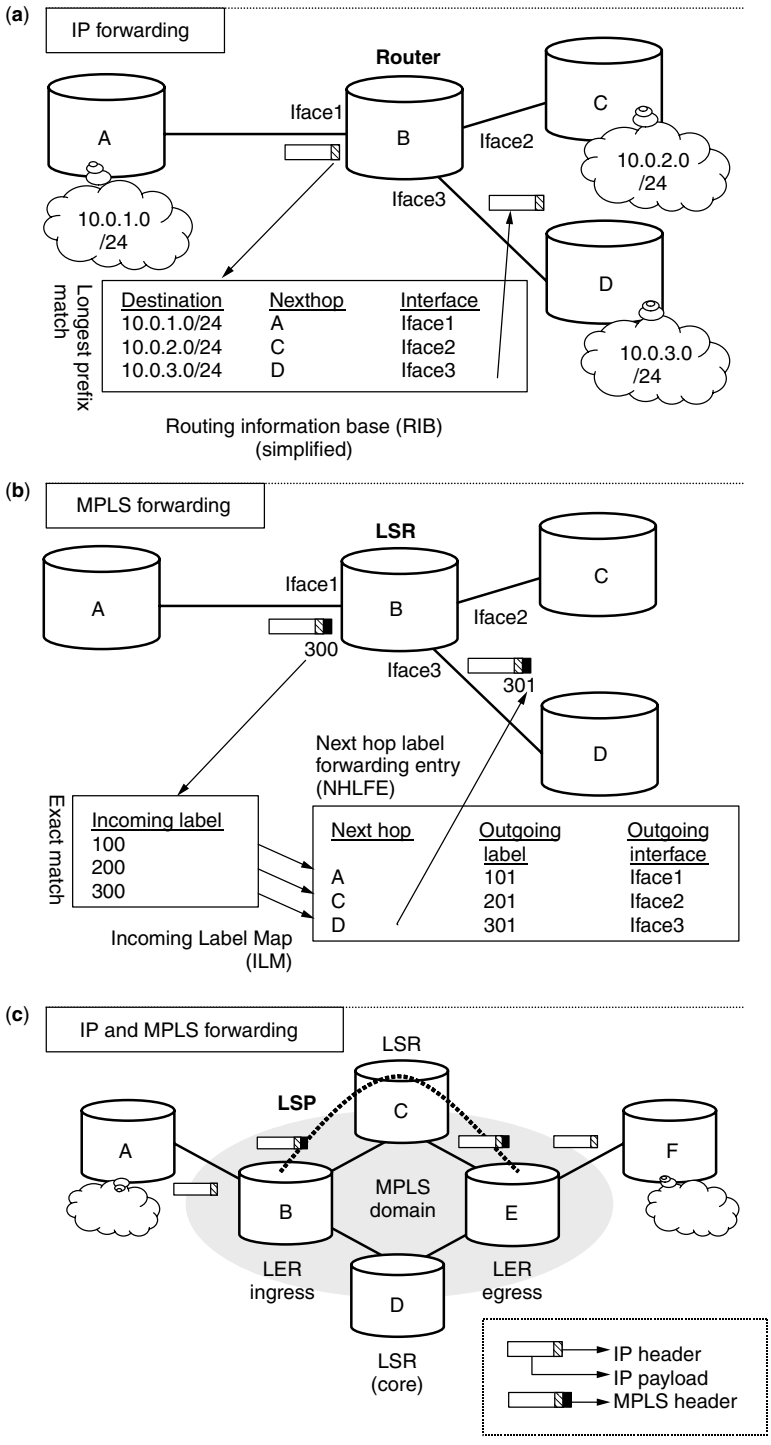


Figure 1. (**a**) IP forwarding for a network consisting of four nodes; (**b**) example of MPLS forwarding; (**c**) IP and MPLS forwarding.

information in the ILM (incoming label) and that in the NHLFE (outgoing label) are combined, effectively storing the information about the label switch (step *e*). After allocating the label and storing the relevant information, LSR *B* will send a label mapping to its previous hop (step *f*). Finally, the initiator of the LSP setup (node *A*) will receive the label mapping from its next hop. LSR *A* will store this information in a NHLFE. This ingress LER will then map traffic to the newly established LSP by mapping a class of packets (FEC) to the LSP, which implies that traffic that belongs to this traffic class will be forwarded over the LSP. The FEC is thus mapped on the NHLFE (step *g*). All the FEC-to-NHLFE mappings are stored in the FEC-to-NHLFE map (FTN). The FTN is used by the ingress of an LSP to forward the packets belonging to a certain FEC over the LSP (to find the outgoing label).

Because the request is forwarded according to the local RIB of the intermediate routers, the resulting LSP is called a *hop-by-hop routed LSP*. Another type of LSP is called an *explicit routed LSP* (ER-LSP).

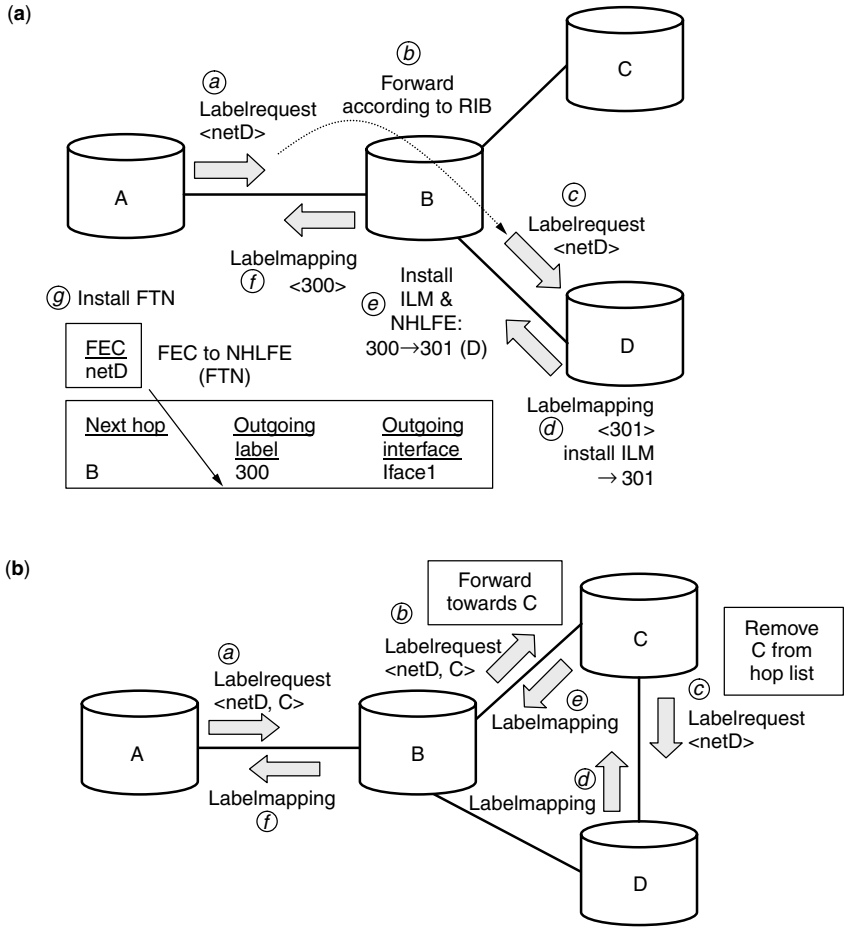


Figure 2. Examples of (**a**) hop-by-hop routed label distribution and (**b**) explicitly routed label distribution.

1.2.2. Explicit Routed LSP Setup. The real power of MPLS lies in the fact that paths can be set up with a great deal of flexibility. An example is an explicit routed LSP (ER-LSP). The term *explicit routed* means that some or all of the hops between the ingress and the egress of the LSP can be specified. Figure 2b illustrates this; in step *a* LSR *A* sends a label request for *netD* and the label request explicitly states that the LSP should be routed along node *C*. Node *B* will receive this label request and will forward it toward *C* along the shortest path (step *b*). When LSR *C* receives this label request, it removes itself from the hop list in the label request and forwards the label request toward the destination. From then on the LSP setup continues as detailed in Fig. 2. It is important to note that every node keeps state about the label request so that the label mappings are sent to the correct previous hop, that is, the hop it received the corresponding label request from.

1.3. Conclusion

In IP both the route calculation and the forwarding is based on the destination address. MPLS separates the forwarding from the route calculation by using labels to forward the packets. To distribute these labels over the domain and hence set up an LSP, MPLS uses a label distribution protocol. LSPs can be setup according to the IP routing tables or the hops to be traversed can be explicitly specified.

2. ARCHITECTURE

After the general overview of MPLS versus IP, this section describes MPLS architecture in greater detail [4].

MPLS architecture consists of a forwarding layer and a signaling layer. The functionality of the forwarding layer is to inspect the incoming label, look up the outgoing label(s), and forward the packet with the new label(s) to the correct outgoing interface (see Section 2.1). The signaling layer is responsible for setup of the MPLS paths (LSPs). The protocols responsible for the setup of LSPs are called *label distribution protocols* (see Section 3.2).

2.1. Forwarding Layer

This section begins with a summary of the most important MPLS forwarding concepts and then gives details on other important issues and terminology with respect to the MPLS forwarding.

2.1.1. MPLS Forwarding Concepts. MPLS architecture formalizes three concepts with respect to the forwarding plane:

1. The *next-hop label forwarding entry* (NHLFE) contains all the information needed in order to forward a packet in a MPLS router. It contains the packet's next hop and the outgoing label

operation. The NHLFE may also contain the data-link encapsulation and information on how to encode the label stack when transmitting the packet.

2. The *incoming label map* (ILM) defines a mapping between an incoming label and one or more NHLFEs. It is used when forwarding labeled packets. If the ILM maps a particular label to more than one NHLFE exactly, one NHLFE must be chosen before the packet is forwarded. Having the ILM map a label to more than one NHLFE can be useful to do, for instance, load balancing over a number of LSPs. Since the ILM is used to forward *labeled* packets in a LSR, it is typically used in a core LSR.
3. Finally, the *FEC-to-NHLFE map* (FTN) maps a FEC to one (or more) NHLFEs. It is used when forwarding packets that arrive unlabeled, but that are to be labeled before being forwarded. The FTN map is used in the ingress *label edge router*.

2.1.2. Label Encapsulation. It is apparent that "labels" constitute the center of MPLS architecture. However, the properties of the labels differ from the link layer on which MPLS is supported. Because of this close tie with the link-layer technology, MPLS is sometimes called a *layer 2.5 architecture*, situated between the link layer (layer 2) and the networking layer (layer 3). Two categories of data-link layers can be distinguished: (1) link layers that natively support fixed-length label entities and switch on them. Examples of [e.g., ATM — see Fig. 3c, virtual circuit identifier (VCI) or virtual path identifier/(VPI), (VPI/VCI) and frame relay — Fig. 3b, data-link circuit identifier (DLCI)] and (2) link-layer technologies that do not natively support labels but encapsulate the labels

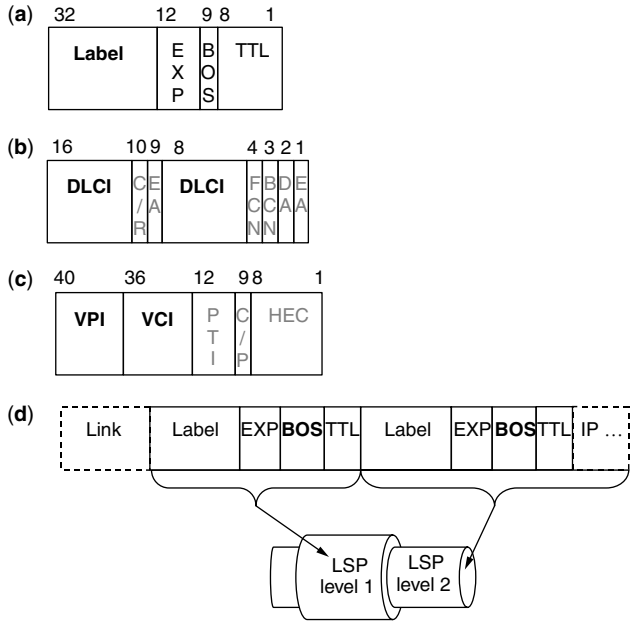


Figure 3. MPLS label encapsulation: (**a**) generic MPLS encapsulation with the shim header; (**b**) encapsulation in the frame relay DLCI; (**c**) encapsulation in the ATM VPI/VCI; (**d**) encapsulation of multiple labels with shim headers (label stacking).

by transmitting an additional header. This small header, called the *shim header*, is inserted between the link-layer header and the networking header. The former way of encapsulating the MPLS labels is called *link-layer-specific encapsulation*, whereas the latter is called *generic MPLS encapsulation*. The shim header contains a label, three experimental bits, a bottom-of-stack (BoS) indicator, and a TTL ("time to live") field (Fig. 3a). The *label* field (20 bits) is used to store the label value, the three experimental (*EXP*) can be used to support Diffserv over MPLS (Diffserv will be covered in detail in Section 2.3.2), and/or early congestion notification (ECN) or other experimental extensions to MPLS forwarding. The *BoS* bit is used to indicate the last shim header of the label stack. Finally, the *TTL* field is used to support the IP time to live mechanism (see Section 2.1.4).

An example of a link-layer technology that has a native label entity is ATM (asynchronous transfer mode). In ATM-based MPLS a label is a VPI, a VPI/VCI, or a VCI identifier. In ATM networks these identifiers are installed with user–network interface (UNI) or private network–node interface (PNNI) signaling. MPLS does not use ATM signaling because a label distribution protocol is used instead. When link-layer-specific label encapsulation is used, the label stack is still encoded in the shim header but the shim header cannot be processed by the intermediate LSRs (only at the ingress and the egress). Therefore the top label of the label stack is copied to the native label entity before the ATM or FR (Frame relay) segment. Similarly after the segment, the current label value from the native label is copied to the top label of the label stack.

2.1.3. Label Operations and Label Stacks. Only a limited number of operations are possible on MPLS labels: (1) replace the top label with a new label (label swap), (2) remove the top label (label pop), or (3) replace the top label and push a number of new labels. This means that multiple labels can be pushed on top of each other, leading to a label stack. This can be useful to aggregate multiple LSPs in one top-level LSP and hence reduce the LSP state. Label stacking is also use in MPLS VPNs (see Section 3.2).

Label operations are possible only on the top label of the label stack. In other words, "pop the label stack" means to remove the top label of the stack. Similarly at every node only the top label is considered when making the forwarding decision.

2.1.4. Support for the IP "Time to Live" (TTL) Field. In regular IP networks, the "time to live" (TTL) field in the IP header is decremented at every hop. When the TTL reaches zero, the packet is dropped. This mechanism is used to prevent packets from being forwarded forever in case of a network anomaly (e.g., a network loop). To support this mechanism in MPLS, the TTL information must be available to the LSR. Since the shim header contains a TTL field, the LSRs are able to decrement the TTL just as in regular IP forwarding.

When MPLS is supported by a link-layer technology that uses its own native label entities, the LSR can act only on these native labels. Unfortunately, these link-layer-specific MPLS labels do not have a TTL field. It

is therefore impossible to decrement the TTL at every hop. The solution is to compute the total TTL decrement for the non-TTL-capable segment (the switching segment) at LSP setup time. When a packet arrives at the first hop of this segment, the precomputed TTL decrement is subtracted from the current TTL. If the result is positive, then the TTL is written to the top entry of the label stack. A packet traveling through the LSP will have the correct TTL both before and after the non-TTL-capable segment. The TTL will have a constant value in the segment (the same value as immediately after the segment). If there is no information about the hop count of the non-TTL-capable segment, the current TTL is decreased by one and the resulting value is written to the label stack. In any case appropriate actions must be taken on this result (e.g., dropping the packet if the TTL reaches zero).

2.1.5. Label Merging. In order to reduce the number of outgoing labels, different incoming labels for the same FEC can use the same outgoing label (label merging).

A link layer is capable of merging labels if more than one incoming label can be mapped to a single outgoing label (all the incoming labels are merged to the same outgoing label). A merge-capable LSR needs to store only the label information of every FEC. However, some link-layer technologies are not capable of merging labels. In this case a one-to-one mapping between incoming and outgoing labels is required. This has scalability implications because an individual LSP state is needed for every ingress–egress pair traversing a given LSR (unless label stacking is be used).

2.1.6. Label Spaces. Another point of difference between some link-layer technologies is the scope of the labels. The MPLS architecture uses the term *label space*. Label spaces limit the scope of a label, which means that the label is valid and defined only for a given label space. Two labels match only if both the label value and the label space are equal. This has the consequence that the same label value can be reused in a different label space. ATM and frame relay have a label space per interface, which means that label A on interface 1 is will be interpreted differently from label A on interface 2. On the other hand, platformwide label spaces are not tied to a specific interface but are valid on the whole platform (i.e., router or host). Global label spaces have the advantage that if the incoming interface of an LSP changes (e.g., during rerouting), no action needs to be taken. Per interface label spaces reduces the number of incoming labels per interface (which might be useful if labels are a scarce resource).

2.1.7. Penultimate-Hop Popping. Since no forwarding decisions have to be made at the last hop of an LSP, the next-to-the-last hop (also referred to as the *penultimate hop*) can also pop the top label of the LSP. The penultimate hop pops the label and sends the packet without a label to the egress node, thereby eliminating the label overhead.

2.2. Signaling Layer

MPLS architecture allows for multiple methods for distributing labels. The reader is referred to Fig. 2 for a basic overview of the label distribution process. The subsequent sections describe the most important label distribution protocols. Some important terminology will be introduced first.

1. *Downstream versus Upstream Allocation*. Since labels only have a local meaning, these labels can be allocated decentralized by the switch controllers. For a given LSP, a core LSR has two neighbors, one upstream and one downstream. So it needs an incoming label and an outgoing label. With that in mind, there are two possible approaches: (1) the LSR supplies its upstream neighbor with a label and receives a label from his downstream neighbor or (2) vice versa. When the LSR receives the label from its downstream neighbor, this is called *downstream allocation*. MPLS typically uses downstream allocation.

2. *Unsolicited Distribution versus Distribution on Demand*. As described in paragraph 1, labels are chosen (allocated) by the downstream LSRs. When these labels are distributed spontaneously (without request), this is called *unsolicited label distribution*. When the upstream LSR always sends a request to its downstream neighbor in order to obtain a label, this is called *distribution on demand*.

3. *Independent versus Ordered Control*. In *independent control*, when an LSR recognizes a particular FEC, it makes an independent decision to map a label to that FEC and to distribute that mapping. In *ordered control*, an LSR maps a label to a particular FEC only if it is the egress for that LSP or if it has already received a label binding from its next hop.

4. *Liberal Retention versus Conservative Retention*.

Consider the situation where an upstream LSR has received and retained a label mapping from its downstream peer. When the routing changes and the original downstream peer is no longer the next hop for the FEC, there are two possible actions the LSR can take: (**a**) release the label, an action called *conservative retention*; or (2) keep the label for later use, which is called *liberal retention*.

Conservative retention (*release on change*) has the advantage that it uses fewer labels; liberal retention (*no release on change*) has the advantage that it allows for faster reaction to routing changes.

5. *Label Use Method*. Labels can be used as soon as a LSR receives them (*use immediate*) or the LSR can only use a certain label if the corresponding LSP contains no loop (*use loop-free*). MPLS supports both loop prevention (preventing an LSP with a loop from being set up) and loop detection mechanisms (detecting whether an LSP contains a loop).

2.2.1. The Label Distribution Protocol (LDP). The *label distribution protocol* (LDP) is the basic signaling protocol proposed by the IETF MPLS Working Group for hop-by-hop routed LSPs [5]. In LDP labels are distributed for a given forward equivalent class (FEC). In LDP a FEC can be either an IP prefix or an IP host address. LDP gives the user a great deal of freedom in how to set up the LSPs. LDP peers use TCP as the transport protocol for LDP messages. This ensures that these

messages are reliably delivered and need not be refreshed periodically (LDP is therefore called a *hard-state protocol*). Session management allows LDP peers (most of the time neighbors) to discover each other and to negotiate about session parameters (e.g., on-demand or unsolicited distribution). After this negotiation phase, a LDP session is set up and the distribution of the labels can start. LSP supports unsolicited and on-demand distribution of labels, liberal and conservative label retention, independent and ordered control, and the immediate or loop-free use of labels (Fig. 2a illustrates ordered control, downstream on-demand label distribution).

Information in LDP messages is encapsulated in type–length–value (TLV) structures. These TLVs are used for standard features but can also be used to extend LDP with experimental and/or vendor-private mechanisms. The constraint-based label distribution protocol (CR-LDP) is an extension to LDP and will be covered in Section 2.2.2.

2.2.2. Constraint-Based Label Distribution Protocol (CR-LDP). CR-LDP introduces a number of extensions to LDP in order to support MPLS traffic engineering (TE). CR-LDP supports only the downstream on-demand ordered label distribution and conservative label retention mode. The additional functionality of CR-LDP compared to LDP is (1) the possibility to setup constraint-based LSPs, (2) the support for traffic parameters, (3) preemption, and (4) resource classes [6].

1. *Constraint-Based Routes*. CR-LDP allows setup of LSPs that defer from the shortest path by explicitly indicating the hops that the LSP should traverse. There's a distinction between strict and loose hops. A "strict" hop on an LSP means that the next hop along the LSP should be that hop and that no additional hops may be present. A "loose" hop on an LSP simply requires that the hop be present on the path that the LSP traverses. Hops are ordered, which means that they should be traversed in the order they are specified. CR-LDP supports the notion of abstract nodes. An *abstract node* is a group of nodes whose internal topology is opaque to the ingress node of the LSP. An abstract node can, for example, be denoted by an IPv4 prefix, an IPv6 prefix, or an autonomous system (AS) number.

2. *Traffic Parameters*. The traffic parameters of an LSP are modeled with a peak and a committed rate. The *peak rate* is the maximum rate at which traffic should be sent over the LSP. The peak rate of an LSP is specified in terms of a token bucket with a rate and a maximum token bucket size. The committed rate is the rate that the MPLS domain commits to the LSP. The committed rate of an LSP is also specified in terms of a token bucket. The extent by which the offered rate exceeds the committed rate may be measured in terms of another token bucket that also operates at the committed rate, but the maximum size of the this token bucket is the maximum excess burst size. There can also be a weight associated with the LSP; this weight indicates the relative share of the available bandwidth the excess bandwidth of an LSP receives. Finally, the frequency of an LSP indicates the granularity at which the committed rate is made available. CR-LDP also provides support for DiffServ over MPLS, as will be described in Section 2.3.2.

3. *Preemption*. An LSP can have two priorities associated with it: a setup priority and a hold priority. The *setup* priority indicates the relative priority an LSP has to use resources (bandwidth) when set up. A LSP with a higher setup priority can be set up in favor of an existing LSP with a lower priority (it can preempt an existing LSP). The *holding* priority indicates how likely it is for an LSP to keep its resources after having been set up.

4. *Resource Classes*. In CR-LDP one can specify which of the resource classes an LSP can traverse. A resource class is usually associated with a link, enabling one to indicate which links are acceptable to be traversed by an LSP. Effectively, this information allows for the network's topology to be pruned; thus, certain links cannot be traversed by the LSP. For example, a provider might want to prevent continental traffic from traversing transcontinental links.

2.2.3. Extensions to RSVP for LSP Tunnels (RSVP-TE). The "extensions to RSVP for LSP tunnels (RSVP-TE)" protocol is an extension of the "resource reservation protocol" [7], a signaling protocol originally developed for Intserv reservations. RSVP-TE First extends RSVP with the possibility to set up LSPs and then adds traffic engineering functionality [8].

2.2.3.1. Setting up Paths with RSVP-TE. RSVP-TE is based on the RSVP protocol, which does not have support to set up LSPs. RSVP-TE has been extended to support this by introducing a new LSP session type and then defining two new objects: (1) a label request object, which is encapsulated in the downstream direction on the RSVP PATH messages; and (2) a label object, which is encapsulated in the upstream direction on the RSVP RESV messages. Labels are allocated in the upstream direction by the downstream nodes. In other words, RSVP implements a downstream on-demand label distribution protocol. RSVP does not have direct support to detect the failure of a neighboring node. To address this, the Hello protocol has been developed. This protocol allows RSVP to detect the liveliness of its neighbors. RSVP also has a loop detection protocol to prevent setting up an LSP that contain loops. This makes RSVP more or less, functionalitywise, equivalent to LDP in downstream on-demand mode, with the important difference that RSVP is a soft-state protocol. This means that the state with respect to the LSP has to be refreshed periodically in the network. The advantage of a soft-state protocol is that the protocol responds more naturally to network changes while it typically requires more signaling overhead.

2.2.3.2. Traffic Engineering with RSVP-TE. Other extensions to RSVP introduce the traffic engineering capabilities very similar to CR-LDP's functionality. Like CR-LDP, RSVP-TE supports the notion of explicitly routed paths whereby the (abstract) hops can be specified strict or loose. The approach to bandwidth and resource allocation differs fundamentally from the CR-LDP model. As mentioned before, RSVP is a signaling protocol for IntServ,

so support for IntServ in RSVP-TE is naturally inherited from the base RSVP protocol. As in CR-LDP, it is possible to indicate the setup and holding priority of the LSP. The resource class procedures for RSVP-TE are more powerful than those found in CR-LDP. In CR-LDP a link is eligible to be traversed by an LSP if the resource class of the link is part of the resource classes specified in the label request message. This leads to the procedure where any link can be used as long as the link is part of the resource class collection specified in the label request message. RSVP-TE also supports this include-any relationship between links and LSPs, but it also supports exclude-any and include-all relationships. It is not necessary to specify any of three relationships, but if set, they must match for the link to be taken into account.

2.2.4. Carrying Label Information in Border Gateway Protocol 4 (BGP4). The Border Gateway Protocol 4 (BGP4) is used to distribute routes across the Internet. These routes can be interdomain routes, making BGP the sole interdomain routing protocol. By piggybacking label information on the BGP route UPDATE messages, BGP can be used to distribute the label mapped to that route. A simple example of the use of BGP as a label distribution protocol is when two BGP peers are directly connected, in which case BGP can be used to distribute labels between them. A more important use of BGP as a label distribution protocol is the more common case where the BGP peers are not directly connected but belong to an MPLS domain that supports another label distribution protocol (e.g., LDP). BGP4 and another label distribution protocol is used to administer MPLS VPNs (see Section 3.2).

2.3. MPLS and Quality of Service

Although MPLS is not a quality-of-service (QoS) framework, it supports delivery of QoS. The following sections describe how the two major models for QoS in IP (IntServ and DiffServ) are implemented with MPLS [9].

2.3.1. Integrated Services. Integrated Services (IntServ) architecture has the goal to provide end-to-end QoS (in the form of services) to applications. The IntServ QoS model has defined two service types: *guaranteed service* (guaranteed delay and bandwidth) and *controlled load* (QoS closely approximating that of an unloaded network). The architecture uses an explicit setup mechanism to reserve resources in routers so that they can provide requested services to certain flows. RSVP is an example of such a setup mechanism, but the IntServ architecture can accommodate other mechanisms. RSVP-TE as an extension of RSVP has natural support for both IntServ service types. An LSP with IntServ reservation is created just like any other IntServ reservation but additionally the MPLS specific LABEL_REQUEST and LABEL objects are piggybacked on the PATH and RESV message, respectively.

CR-LDP does not have support for IntServ natively, but it can support (a number of) IntServ flows over an LSP by setting the appropriate traffic parameters of the LSP. In order to guarantee the service received on the LSP, admission control and policing on the ingress is required.

2.3.2. Differentiated Services. In order to solve the IntServ scalability problem, Differentiated Services (DiffServ) classifies packets into a limited number of *classes* and therefore does not need for per flow state or per flow processing. The identified traffic is assigned a value, a DiffServ code point (DSCP). A DiffServ *behavior aggregate* (BA) is a collection of packets with the same DiffServ codepoint (DSCP) crossing a link in a particular direction. A per hop behavior (PHB), the externally observable forwarding behavior, is applied to a behavior aggregate.

The classification is usually based on multiple fields in the IP header (multifield, MF classification) at the edge and on the DiffServ codepoint (behavior aggregate, BA classification) in the core of the network (see Fig. 4c). An example PHB is *expedited forwarding* (EF), which offers low loss, low delay, and low jitter with an assured bandwidth. This means that the collection of packets marked with the EF codepoint traversing a link in a certain direction (BA) will receive low loss, delay, jitter, and an assured bandwidth. The *assured forwarding* (AF) PHB group is a group of PHB. A PHB of the AF group is denoted as *AFxy*, where x is the class and y is the drop precedence. Packets belonging to a different AF class are forwarded separately. Usually more resources are allocated to the lower classes. Packets within a class that have a higher drop precedence will be dropped before packets with a lower drop precedence.

An *ordered aggregate* (OA) is the set of behavior aggregates that share an ordering constraint. This means that packets that belong to the same OA must not be reordered. When looking at DiffServ over MPLS, it immediately becomes apparent that packets that belong to a certain OA must be mapped on the same LSP; otherwise this ordering constraint cannot be enforced. This is trivial if only one PHB is applied to the ordered aggregate. However, PHBs can be grouped in a per hop behavior scheduling class (PSC). A PSC is the set of one or more PHB(s) that are applied to a given OA. For example, AF1y is a PSC comprising the AF11, AF12, and AF13 PHBs. Combining the notion of OA and PSC means that in DiffServ over MPLS, OA-PSC pairs will be mapped on LSPs. If the PSC contains more than one PHB, this means that it must be possible for an LSR to enforce different PHBs to the packets that belong to the same LSP. This, in turn, means that the LSR must have some information in the packet header to determine the PHB to be applied. This information must be encapsulated in a way that is accessible to the LSR and thus must be part of the label or shim header. We will now discuss how to encapsulate the PHB.

2.3.2.1. Encapsulating the PHB. In IPv4 the "type of service" (ToS) field or in IPv6 "traffic class" field is used to encapsulate the DSCP. With generic MPLS encapsulation there is a mapping from the IP DSCP space to the EXP field of the shim header [10]. The DSCP field uses the 6 most significant bits of the 8 bits of these IP header fields. Since the DSCP field is 6 bits wide, it can represent 64 different values. However, the EXP field of the shim header is only 3 bits wide, so it can represent only 8 different values. This means that the mapping from DSCP to EXP value cannot be a one-to-one mapping. This is quite a problem because currently there are more than eight defined DSCP values

(best effort, 12 AF values, and EF). If the DiffServ domain uses less than 8 different DSCP values, then the mapping between DSCP and EXP can be fixed over the domain. If the domain uses more than eight different codepoints, then the mapping must be explicitly defined on a per LSP basis.

When the EXP value is used to indicate the set of PHBs applied to an OA (the PSC), we call this an *EXP-inferred-PSC LSP* (E-LSP). This means that the PSC is inferred from the EXP value in the shim header (see Fig. 4b).

This works only for LSRs that support shim headers but link layer specific labels do not have an EXP field. The solution is to set up a distinct LSP for each FEC and ordered aggregate (FEC-OA) pair and signal the PSC during the LSP setup. When the PSC of an OA contains more than one PHB, these different PHBs still need to be enforced. The PHBs of the PSC differ only in drop precedence; thus we need to encapsulate the drop precedence in the link-layer specific label. In ATM only the cell loss priority (CLP) bit can be used to encapsulate this information. Similarly, the discard eligibility (DE) bit of frame relay can be used to encapsulated the drop precedence (shown in Fig. 3). An LSP where the PSC is inferred from the label value is called a *label-only-inferred-PSC LSP* (L-LSPs), meaning that the PSC is inferred from the label values as opposed to the EXP field value (see Fig. 4c).

The use of L-LSPs is not restricted to link-layer-specific label encapsulating LSRs; it can also be used with generic MPLS encapsulation. The drop precedence is then encapsulated in the EXP field of the shim header, and the PSC is still inferred from the label value.

2.3.2.2. Allocation of Bandwidth to L-LSP and E-LSPs. Bandwidth can be allocated to E-LSP and L-LSPs at setup time. When resources are allocated to an L-LSP, the bandwidth is allocated to the PSC of the LSP; when bandwidth is allocated to an E-LSP, then the bandwidth is associated to the whole LSP, that is, the set of PSCs of the LSP. Signaling bandwidth requirements of the LSPs can be useful in two ways: (1) associating bandwidth to an LSP can be used to admit the traffic to the LSP according to the availability of resources and (2), the bandwidth allocation information can also be used to shift resources from certain PSCs to others. It is important to note that allocating resources to an L-LSP or E-LSP does not lead to the necessity of having a per LSP forwarding treatment.

2.4. History

MPLS started out as a technique for IP over ATM interworking as a convergence of a number of "IP

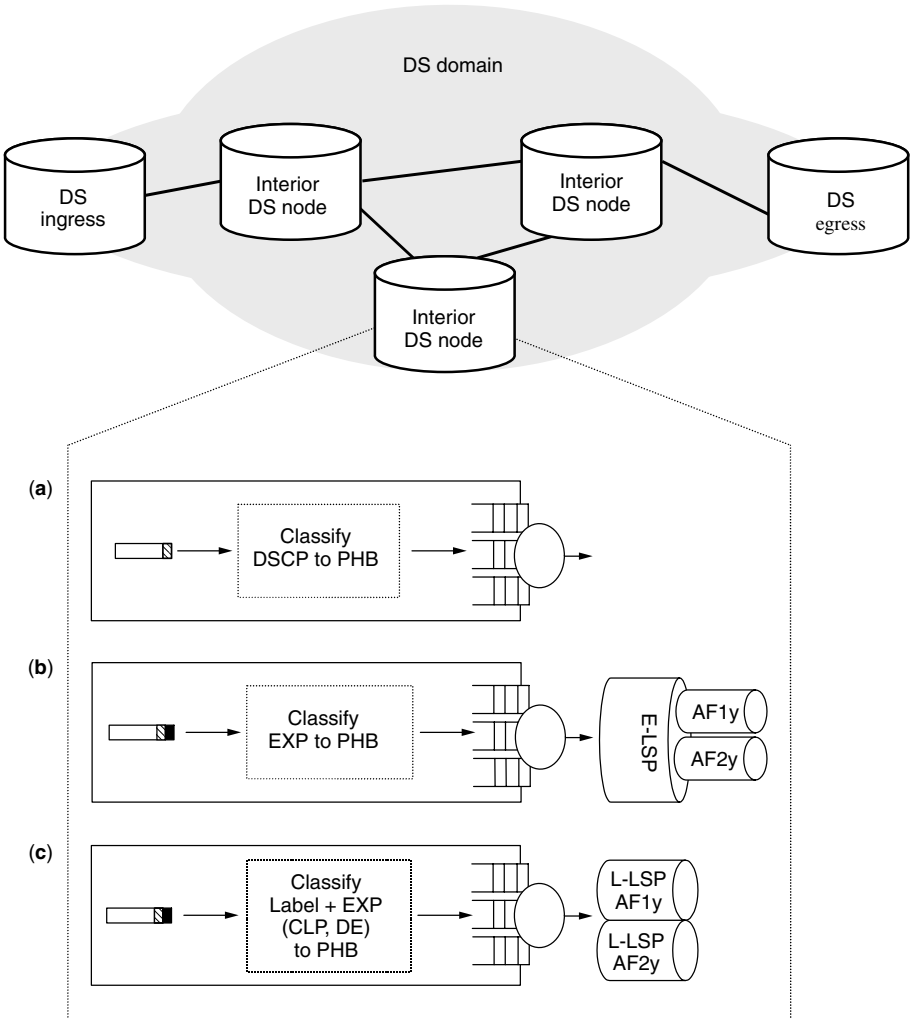


Figure 4. BA classification in DiffServ: (**a**) classification on the DSCP value of the IP header; (**b**) classification on the EXP bits of the shim header; (**c**) classification on the label and the EXP bits (or the ATM CLP or FR DE bit).

switching" schemes. IP switching is a technique that uses ATM hardware to forward of IP packets. In contrast to ATM networks, in MPLS networks the ATM hardware is administrated by IP and MPLS signaling protocols, and not by ATM signaling. There are a number of different IP switching implementations: Cisco Systems Tag Switching, IBM's Aggregated Route-based IP Switching (ARIS), Toshiba's Cell Switch Router (CSR), and NEC's Ipsofacto [11]. In order to standardize all these IP switching techniques, a new IETF work group came to life in 1997. The MPLS work group has since then been working on forming a common technology for IP switching.

MPLS contrasts with a number of other techniques for IP over ATM, which are overlay techniques. Examples of overlay techniques are multiprotocol over ATM (MPOA) and the work done in the IETF Internetworking Over Non-broadcast-capable networks (ION) working group. In the case of an ATM overlay network, there are two distinct networks: an ATM network and an IP network. This leads to disadvantage of having to administer the two networks and the fact that the scalability is limited due to full meshed peering [12].

3. APPLICATIONS OF MPLS

The following section will describe arguably the three most important applications of MPLS: traffic engineering, virtual private networks, and resilience (more specifically, fast rerouting).

3.1. Traffic Engineering

Traffic engineering (TE) is generally defined as the performance optimization of operational networks. (Other definitions focus more on the role of traffic engineering to offer efficient services to customers.) While TE is not strictly tied to multiservice networks and QoS, it is definitively more complex and mission-critical in multiservice networks. TE is probably considered as the most important application of MPLS networks.

In the following sections we will first address the applicability of TE and then discuss how these techniques are implemented in MPLS and how they relate to regular IP implementations.

3.1.1. Applicability. The optimization of operational networks is typically achieved by (1) the avoidance of congested routes, the (2) resource utilization of parallel links, and (3) routing policies using affinities.

1. *Avoiding Congested Routes*. When certain network segments are congested while others are underutilized, the network operator will want to route traffic away from the congested segments. In regular IP this can be done by modifying static link metrics [13] or using dynamic metrics, but this is difficult because of the destination-based forwarding of IP. In MPLS traffic can be routed away from the congested segments by setting up (explicit routed) LSPs. Traffic can be mapped on an LSP not only according to the destination, but virtually any classification can be used. A small number of LSPs can be used to route traffic away from the congested segments or a mesh of LSPs can be set up to distribute the traffic evenly over the network.

2. *Resource Utilization of Parallel Paths*. Regular IP calculates and uses only a single shortest path from point A to point B. This limitation is addressed by equal-cost multipath (ECMP) extensions to routing that takes paths of equal cost into account and spreads the traffic evenly over the available paths with the same cost. Even more advanced is the optimal multipath (OMP) extension, where paths with different cost values are used and the traffic is spread according to the relative cost (e.g., a path with a higher cost gets a lower share of the traffic). The cost metric of OMP can be dynamic; that is, it can be based on the actual load and length of the path. MPLS can be used to explicitly configure parallel paths. The calculation can be based on the online (routing) mechanisms such as ECMP or OMP, or alternatively an offline TE algorithm can compute the paths. Offline LSP calculations can be based on the measured and forecasted traffic between the edge nodes of the networks (the traffic matrix).

3. *Routing Policies*. A network operator might want to exclude some types of traffic from certain links or force traffic on certain links. In MPLS this can be achieved by using the resource class procedures of RSVP-TE or CR-LDP. In IP traffic engineering, extensions for OSPF or IS-IS have been defined to cope with resource affinity procedures.

3.1.2. Implementation. MPLS traffic engineering allows one to gain more network efficiency. But there's no such thing as a free lunch. Efficiency can be gained by introducing more LSPs in the network, but there's a tradeoff between the control granularity and the operational complexity associated with a large number of LSPs. In order for the traffic engineering to work properly, it is necessary to obtain detailed information about the behavior of LSPs (LSP monitoring). This is not a trivial task and can require significant resources. The assignment of resources to LSPs and the mapping of traffic to LSPs is another task that can be both time-consuming (for the network operator or in terms of computing power) and prone to errors due to inaccurate or outdated traffic matrices. Path calculation is another additional task to perform in comparison with non-traffic engineered networks. Finally, the signaling overhead introduced by traffic engineering can cause additional overhead.

An alternative for a traffic-engineered network is an overprovisioned network that always has enough capacity to transport the offer load. Even in overprovisioned networks, monitoring is necessary to determine when to upgrade the network capacity.

3.2. Virtual Private Networks

A *virtual private network* (VPN) is a network using secure links over the public IP infrastructure [14]. A VPN is a more cost-effective solution to a corporate extranet than a private network, which consists of private infrastructure. In order to create the extranet, the different sites have to be interconnected through the provider's network (ISP network). The access points between a customer's site and the provider are called *customer edge* (CE) and *provider edge* (PE), respectively. The internal routers are called *provider* (P) *routers* (see Fig. 5). A VPN consists of number

of sites that are connected through the ISP network. A participating site can be part of more than one VPN (e.g., site CE4). If two VPNs have no sites in common, then the VPNs can have overlapping address spaces. Since the addresses can overlap, the routers need to interpret the addresses on a persite basis by installing persite forwarding tables in the PE routers. MPLS VPNs are set up with the combination of BGP4 and another label distribution protocol (LDP, CR-LDP, or RSVP-TE) (see Section 2.2.4). The PE routers distribute labels associated with VPN routes to each other with BGP4. A VPN route is the combination of an IP prefix and a *router distinguisher* (RD). The RD allows one to distinguish between common prefixes of the different VPNs. The other label distribution protocol is used to create a mesh of LSPs between the PE routers. The VPN route labels and the labels distributed by the internal label distribution protocol are used by the PE routers to forward packets over the VPN. The internal LSRs (P routers) operate only on the top label, the label distributed by the internal label distribution protocol, so they don't need to be aware of the BGP routes.

Consider the example of a packet that needs to be forwarded from CE2 toward CE4 over VPN1. The ISP network consists of BGP peers on the edge (PE1 and PE2) and interior LSRs (P1 and P2). A BGP node sends a packet to a certain VPN by looking up the label it has received from his BGP next hop and pushes this label on the label stack. For example, PE1 pushes the label it has received from PE2 for VPN1. PE1 then looks up the label received from the internal label distribution protocol to PE2 and pushes this label on the label stack. The label stack then contains the BGP label for the VPN route (the VPN label) and on top of that the label for the BGP next hop (PE2). Then regular label switching is used to forward the packet to PE2. At PE2 the top label is popped and the VPN label pushed by PE1 is revealed. PE2 will then use this VPN label to look up the information needed to forward the packet to the next hop on VPN1, namely, CE4.

3.3. Resilience

Regular IP typically recovers from network (node and link) failures by rerouting. The routing protocol that is used to calculate the shortest paths in a network is able to detect network failures (or is notified of) and takes them into account when finding new routes after the failure. It typically takes some time before the routing protocol converges, that is, before the network reaches a stable state after the failure.

When using MPLS, IP routing can also be used to restore the shortest-path-routed LSPs. This rerouting in MPLS depends on the new paths calculated by the IP routing protocol, this means that MPLS rerouting based on IP rerouting is slower than IP rerouting.

However, MPLS allows the use of more advanced resilience schemes. *Protection switching* is a scheme where a recovery path is preestablished [15]. The recovery path can be node or link disjunct from the working path. (The working path is used as long as no failures are detected.) When a failure occurs, the traffic is switched from the working path to the recovery path (the protection switch). The use of protection switching leads to a much lower convergence time. An additional advantage of protection switching over rerouting is that resources can be allocated in advance so that even after a failure the traffic over the LSP can still be serviced according to the predefined traffic parameters. Rerouting typically does not offer such guarantees unless the network is carefully planned. The drawback of protection switching is that it requires recovery paths that are preestablished, leading to administrative and signaling overhead and a higher resource usage if the resources are dedicated (i.e., cannot be used when the recovery path is not in use).

BIOGRAPHIES

Pim Van Heuven graduated in computer science from Ghent University in 1998. At this same university

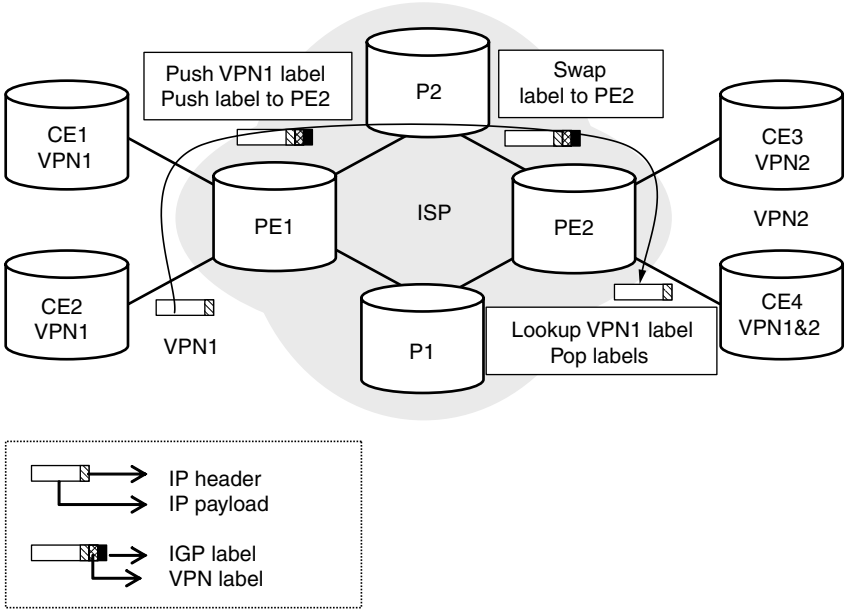


Figure 5. Example of an MPLS VPN and the forwarding of packets between the different interconnected sites.

he joined (August 1998) the Integrated Broadband Communications Networks Group (IBCN), where is now preparing a Ph.D. In January 1999 he was granted an IWT scholarship. He worked on the ACTS IthACI project. Since begin 2000 he has been working on the IST Tequila project. His research interests include MPLS, network resilience, and the areas of quality of service and traffic engineering. Since 2001, he has led the "RSVP-TE daemon for DiffServ over MPLS under Linux" project. He has published several papers on network resilience in IP, MPLS, and G-MPLS.

Steven Van den Berghe graduated in computer science from Ghent University in 1999. In July 1999, he joined the Broadband Communications Networks Group (IBCN) and now is preparing a Ph.D. In January 2001 he was granted an IWT scholarship. His research interests include mainly the areas of quality of service and traffic engineering in IP. He is focusing on measurement-based traffic engineering in a Diffserv/MPLS/multipath environment. He is active in the IST Tequila project and in development of DiffServ support for MPLS in the Linux community and has published, in addition to several papers, an Internet draft on the requirements for measurement architectures for use in traffic-engineered IP networks.

Filip De Turck received his M.Sc. degree in electronic engineering from the Ghent University, Belgium in June 1997. In May 2002, he obtained the Ph.D. degree in electronic engineering from the same university. From October 1997 to September 2001, Filip De Turck was Research Assistant with the Fund for Scientific Research — Flanders, Belgium (FWO-V). At the moment, he is affiliated with the Broadband Communications Networks Group (IBCN) of Ghent University as a Postdoctoral Researcher. His research interests include scalable software architectures for telecommunication networks and service management, performance evaluation and optimization of routing, admission control, and traffic management in telecommunication systems. He is the author of several research papers in this area and has served as a technical program committee member for several international conferences.

Piet Demeester (Senior Member IEEE) received his Ph.D. degree from Ghent University in the Department of Information Technology (INTEC) in 1988. He became Professor at the Ghent University, where he is currently responsible for the research on communication networks. He was involved in several European COST, ESPRIT, RACE, ACTS, and IST projects. He is a member of the editorial board of several international journals and has been a member of several technical program committees (ECOC, OFC, DRCN, ICCCN, IZS, etc.). His current interests are related to broadband communication networks (IP, MPLS, ATM, SDH, WDM, access, active, mobile) and include network planning, network and service management, telecom software, internetworking, and network protocols for QoS support. He has published over 200 papers in this field.

BIBLIOGRAPHY

1. C. Huitema, *Routing in the Internet*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
2. U. Black, *MPLS and Label Switching Networks*, Prentice-Hall, Englewood Cliffs, NJ, 2001.
3. B. Davie and Y. Rekhter, *MPLS Technology and Applications*, Morgan Kauffman, San Francisco, 2000.
4. E. Rosen, A. Viswanathan, and R. Callon, *Multiprotocol Label Switching Architecture*, IETF RFC (Online), *http://www.ietf.org/rfc/rfc3031.txt* (Jan. 2001).
5. L. Andersson et al., *LDP Specification*, IETF RFC (online), *http://www.ietf.org/rfc/rfc3036.txt* (Jan. 2001).
6. B. Jamoussi et al., *Constraint-Based LSP Setup Using LDP*, IETF RFC3212 (online), *http://www.ietf.org/rfc/rfc3212.txt* (Jan. 2002).
7. D. Durham and R. Yavatkar, *Inside the Internet's Resource reSerVation Protocol*, Wiley, New York, 1999.
8. D. Awduche et al., *RSVP-TE: Extensions to RSVP for LSP Tunnels*, IETF RFC (online), *http://www.ietf.org/rfc/rfc-3209.txt* (Dec. 2001).
9. Z. Wang, *Internet QoS: Architectures and Mechanisms for Quality of Service*, Morgan Kaufmann, San Francisco, 2001.
10. F. Le Faucheur et al., *Multi-Protocol Label Switching* (*MPLS*) *Support of Differentiated Services* (online), *http://www.ietf.org/rfc/rfc3270.txt* (May 2002).
11. I. Andrikopoulos et al., Experiments and enhancement for IP and ATM integration: The IthACI project, *IEEE Commun. Mag.* **39**(5): 146–155 (2001).
12. G. Armitage, MPLS: The magic behind the myths, *IEEE commun. Mag.* **38**(1): 124–131 (2000).
13. B. Fortz and M. Thorup, Internet traffic engineering by optimizing OSPF weights, *Proc. IEEE INFOCOM 2000*, Vol. 1, 2000, pp. 519–528.
14. J. Guichard and I. Pepelnjak, *MPLS and VPN Architectures: A Practical Guide to Understanding, Designing and Deploying MPLS and MPLS-Enabled VPNs*, Cisco Press, Indianapolis, 2000.
15. P. Van Heuven et al., Recovery in IP based networks using MPLS, *Proc. IEEE Workshop on IP-oriented Operations & Management IPOM'2000*, IEEE, 2000, pp. 70–78.

MULTIUSER WIRELESS COMMUNICATION SYSTEMS

ASHUTOSH SABHARWAL
BEHNAAM AAZHANG
Rice University
Houston, Texas

1. INTRODUCTION

The 1980s and 1990s witnessed the rapid growth and widespread success of wireless connectivity. The success of wireless systems is due largely to breakthroughs in communication theory and progress in the design of low-cost power-efficient mobile devices. Beyond the widespread use of voice telephony, new technologies are replacing wires

in virtually all modes of communication. For example, in addition to widely recognized outdoor connectivity via cellular wide-area networks (WANs), wireless local-area networks (LANs) and wireless personal-area networks (PANs) have also become popular. Wireless LANs (e.g., IEEE 802.11) provide high-speed untethered access inside buildings replacing traditional wired Ethernet, and wireless PANs (e.g., Bluetooth) are replacement for wires between common peripherals like mouse, keyboard, PDAs, and printers.

Providing ubiquitous mobile access to a large number of users requires solution to a wide spectrum of scientific and economic issues, ranging from low-power semiconductor design and advanced signal processing algorithms to the design and deployment of large cellular networks. In this article, we will highlight the challenges in the design of advanced signal processing algorithms for high-speed outdoor cellular access. The signal processing algorithms form the core of all wireless systems, and are thus critical for their success. In addition, the techniques and algorithms discussed here form a basis for most wireless systems, and thus have a wider applicability than outdoor wireless systems. To keep the discussion tractable, we will focus on baseband design for third-generation wireless cellular systems (e.g., WCDMA or CDMA2000) based on code-division multiple access (CDMA).

A wireless channel is a shared resource; multiple users in the same geographic locale have to contend for the common spectral resource and in the process interfere with other users. To allow meaningful and resource-efficient communication between different users, it is crucial that all participating users agree on a common protocol. The common protocol should enable fair access to the shared resource for all users. The three most commonly used multiple access protocols[1] are time-division (TDMA), frequency-division (FDMA), and code-division multiple access (CDMA). Among the three, direct-sequence CDMA (DS-CDMA) has been adopted as the access technique for all the third-generation wireless standards, and thus is the main focus of this article.

In outdoor cellular systems, the coverage area is divided into smaller regions called *cells*, each capable of supporting a subset of the users subscribing to the cellular system. The cellular structure exploits the fact that electromagnetic signals suffer loss in power with distance, thereby allowing reuse of the same communication channel at another spatially separated location. The reuse of communication channels allows a cellular system to support many more users as compared to a system that treats the whole geographic region as one cell. Each cell is served by a *base station* that is responsible for operations within a cell, primarily serving calls to and from users located in the respective cell. Figure 1 shows the components of a typical cellular system. The size and distribution of the cells [1] are dictated by the coverage area of the base station, subscriber density, and projected demand within a geographic region. As mobile users travel from cell to cell, their calls are *handed off* between cells in order to maintain seamless service. The base stations are connected to the *mobile telephone switching office* (MTSO), which serves as a controller to a group of base stations and as an interface with the fixed wired backbone.

Wireless networks, like typical multiple access networks, have a layered architecture [2,3]. The three main layers of each network are the physical layer, the network layer,[2] and the application layer. The physical layer is responsible for actual transport of data between the source and the destination points. The network layer controls the communication session, and the user applications

[1] We limit our discussion to circuit-switched networks and deterministic multiple-access schemes. In packet-switched networks, probabilistic multiple access is used; a good example is contention avoidance/resolution based protocol used in IEEE 802.11, and packet services used in EGPRS and 3G systems.

[2] The network layer consists of several layers that include the multiple-access layer (MAC), the data-link layer, and the transport layer.

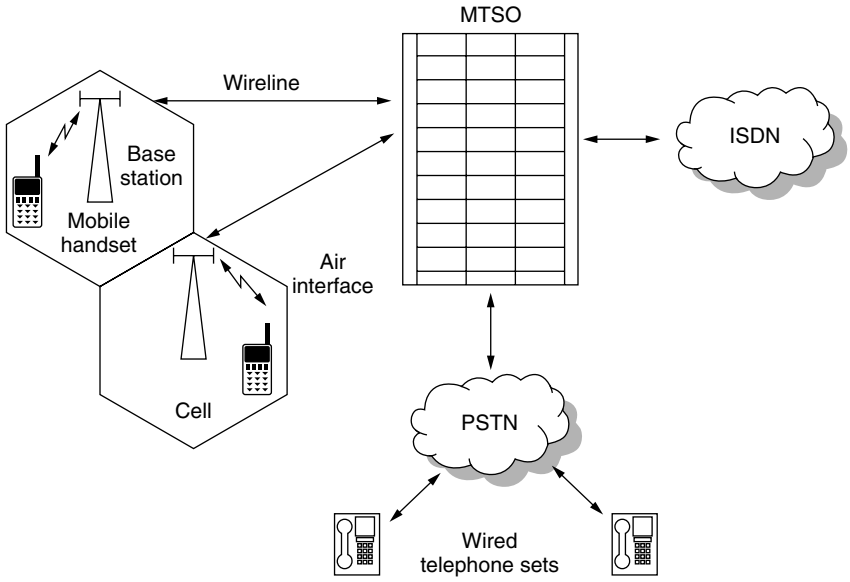


Figure 1. Components of a cellular wireless network.

operate in the application layer. Both network and application layer designs are critical in wireless networks, and are areas of active research. In this article, our focus will be on the design of physical layer for wireless networks.

The rest of the article is organized as follows. In Section 2, we will briefly discuss the three major challenges in the design of wireless systems and commonly used methods to combat them. Models for wireless channels are discussed in Section 3. In Section 4, we will introduce information-theoretic methods to analyze the limits of wireless systems. The core of the article is in Section 5, which discusses various aspects in the design of a typical transceiver. We conclude in Section 6.

2. CHALLENGES AND DESIGN OF WIRELESS SYSTEMS

In this section, we highlight the major challenges and techniques employed in wireless system design.

2.1. Time-Varying Multipath

Enabling mobility, which is the fundamental premise in designing wireless systems and is the major reason for their success, also presents itself as the most fundamental challenge. Because of the mobility of users and their surrounding environment, wireless channels are generally time-varying. Electromagnetic signals transmitted by base station or mobile users reach the intended receiver via several paths; the multiple paths are caused by reflections from man-made and natural objects (Fig. 2). Since the length of each path may be different, the resultant received signal shows wide fluctuations in its power profile (Fig. 3), thereby complicating the design of spectrally efficient systems.

To combat time-varying fading, a combination of time, spatial or frequency diversity is commonly used [4]. By using diversity techniques, the receiver obtains multiple copies of the transmitted signal, thereby increasing the chance that at least one of the copies is reliable. To exploit time diversity, error control codes are used in conjunction with an interleaver [4]. Spatial diversity can be obtained by using multiple antennas that are sufficiently separated.

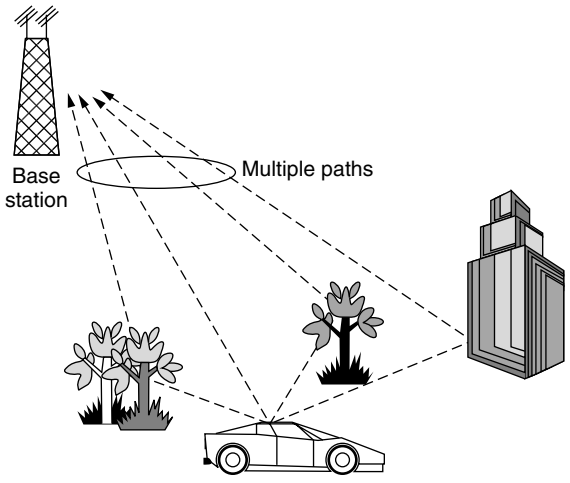


Figure 2. Multipath propagation.

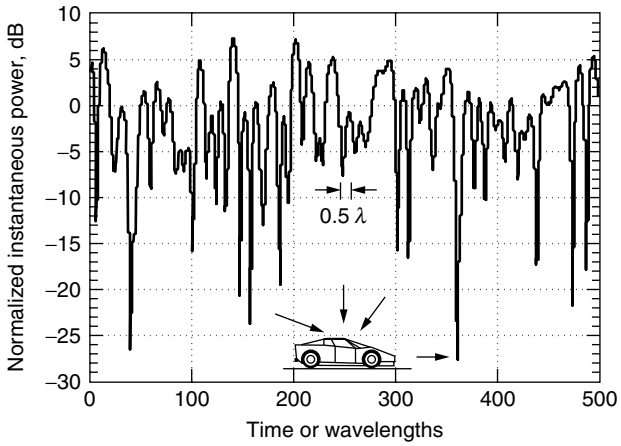


Figure 3. Time variations in the received signal power due to multipath and user mobility.

Spatial diversity can be tapped by using space-time codes [5] at the transmitter or signal combining [6] at the receiver. Spatial diversity techniques have received considerable interest because of their potential to support larger data rates on the same channels compared to current technology. Frequency diversity is analogous to spatial diversity where frequency selectivity due to multipath is used.

2.2. Shared Multiple Access

Unlike wired networks, where new bandwidth is "created" by adding additional physical resources (cables, servers, etc.), users in wireless systems have to share limited spectral resources. Although, the available spectrum for commercial wireless system has increased since 1980, it is clear that growth in demand will always outpace the available spectrum. Limited growth of resources immediately implies that the requirements of new data rate hungry wireless services can be sustained only by progress in efficiently using the available spectrum. An obvious way of increasing system capacity is to use smaller cells, but using smaller cells is undesirable for economic reasons; an increased number of base stations and the required wired backbone are the major reasons for the increased system cost. Further, smaller cells generally lead to increased intercell handoffs and out-of-cell interference, leading to diminishing returns with increasing cell partitioning.

The capacity of cellular systems can also be improved by cell sectorization [7,8], where each cell is further divided into sectors. Cell sectorization is achieved by using multiple directional antenna [9] at each base station, thereby reducing the intersector interference. Because of the directional antenna response, cell sectorization has also been shown to reduce the delay spread of the received signal leading to power savings [10]. Much like cell splitting, cell sectorization also has its limits. To achieve smaller sectors using directional antennas requires increasingly large antennas, which are both expensive and hard to deploy.

Information-theoretic results [11] for multiuser systems indicate that the optimal methods to share spectral

resources should not attempt to avoid intercell and intracell interference. The cochannel interference in wireless systems can be suppressed by using multiuser detection [12], leading to increased spectral efficiency [13,14]. Further improvements in system capacity can be obtained by the use of dynamic resource allocation among users, for example, adaptive channel assignment techniques [15] and dynamic spreading gain and power control [16].

2.3. Power Limitation for Mobile Users

Since most of the mobile devices are battery-operated, power efficiency is a crucial design parameter in wireless systems. The major consumers of power in wireless handsets are power amplifier used during transmission, silicon based computing units (A/D, D/A, and baseband processor) used in reception, and in some cases, the color display.

Power dissipation in the RF power amplifier can be reduced by using cells with smaller radii, better multiuser signal processing at the base-station, improved coding schemes, or receiver diversity. As pointed out earlier, cell splitting is not attractive because of increased system cost with diminishing returns. Advanced signal processing, multiuser channel estimation, and data detection have been shown to greatly reduce the power requirements to achieve a desired performance level [12]. More recent advances in channel coding, namely, Turbo coding [17], can lead to further reduction in power requirements for the transmitter to achieve a desired performance level. Reduction in power requirements of baseband processing units requires development of hardware-frugal algorithms and low-power CMOS circuits. Also, techniques that require more computation at the base-station to cut the complexity of handset are very effective in saving power at the mobile unit.

3. FADING CHANNEL MODELS

In this section, we will describe time-varying wireless channels and the statistical models used to capture their effect on transmitted signals. A detailed discussion of channel models can be found elsewhere [4,18]. A fading multipath channel is generally modeled as a linear system with time-varying impulse response[3] $h(t;\tau)$. The time-varying impulse response is assumed to be a wide-sense stationary random process with respect to the time variable t. Because of time variations of the channel, the transmitted signal is spread in frequency; the frequency spreading is called *Doppler spreading*. The transmitted signal also suffers time spreading as a result of multipath propagation. Thus, the received signal is spread both in time and frequency.

Two parameters are commonly used to characterize wide-sense stationary channels: *multipath delay spread* and *Doppler spread*. To define the multipath delay and Doppler spread, it is convenient to work with the scattering function $\mathcal{H}(\tau;\lambda)$, which is a measure of average power output[4] of the channel at delay τ and frequency offset λ relative to the carrier. The *delay power spectrum* of the channel is obtained by averaging $\mathcal{H}(\tau;\lambda)$ over λ:

$$\mathcal{H}_c(\tau) = \int_{-\infty}^{\infty} \mathcal{H}(\tau;\lambda)\, d\lambda \tag{1}$$

The multipath delay spread T_m is the maximum delay τ for which the delay power spectrum $\mathcal{H}_c(\tau)$ is nonzero. Similarly, the Doppler spread B_d is the maximum value of λ for which the following *Doppler power spectrum* $\mathcal{H}_c(\lambda)$ is nonzero:

$$\mathcal{H}_c(\lambda) = \int_{-\infty}^{\infty} \mathcal{H}(\tau;\lambda)\, d\tau \tag{2}$$

The reciprocal of the multipath delay spread is defined as *channel coherence bandwidth*, $B_{\text{coh}} = 1/T_m$ and provides an indication of the width of band of frequencies that are similarly affected by the channel. The Doppler spread provides a measure of how fast the channel variations are in time. The reciprocal of Doppler spread is called *channel coherence time* $T_{\text{coh}} = 1/B_d$. A large value of T_{coh} represents a slowly fading channel and a small values represents fast fading. If $T_m B_d < 1$, then the channel is said to be *underspread*; otherwise it is *overspread*. In general, if $T_m B_d \ll 1$, then the channel can be accurately measured at the receiver, which can aid in improving the transmission schemes. On the other hand, channel measurement is unreliable for the case of $T_m B_d > 1$.

An appropriate model for a given channel also depends on the transmitted signal bandwidth. If $s(t)$ is the transmitted signal with the Fourier transform $S(f)$, the received baseband signal, with the additive noise, is

$$\begin{aligned} z(t) &= \int_{-\infty}^{\infty} h(t;\tau)s(t-\tau)d\tau + \nu(t) \\ &= \int_{-\infty}^{\infty} H(t;f)S(f)e^{j2\pi ft}df + \nu(t) \end{aligned}$$

where $H(t;f)$ is the Fourier transform of $h(t;\tau)$ with respect to τ. If the bandwidth W of the transmitted signal $S(f)$ is much smaller than the coherence bandwidth, $W \ll B_{\text{coh}}$, then all the frequency components in $S(f)$ undergo the same attenuation and phase shift during propagation. This implies that within the bandwidth of the signal, the transfer function $H(t;f)$ is constant in f, leading to a *frequency nonselective* or *flat fading*. Thus, the received signal can be rewritten as

$$\begin{aligned} z(t) &= H(t;0)\int_{-\infty}^{\infty} S(f)e^{j2\pi ft}df + \nu(t) \\ &= H(t)s(t) + \nu(t) \end{aligned} \tag{3}$$

where $H(t) \in \mathbb{C}$ is the complex multiplicative channel. A flat fading channel is said to be *slowly fading* if the symbol

[3] A linear time-invariant system requires a single-variable transfer function. For a time-varying linear system, two parameters are needed; the parameter t in $h(t;\tau)$ captures the time-variability of the channel.

[4] Under the assumption that all different delayed paths propagating through the channel are uncorrelated.

time duration of the transmitted signal T_s is much smaller than the coherence time of the channel, $T_s \ll T_{\text{coh}}$. The channel is labeled as *fast fading* if $T_s \geq T_{\text{coh}}$.

If the signal bandwidth W is much greater than the coherence bandwidth of the channel, then the frequency components of $S(f)$ with frequency separation more than B_{coh} are subjected to different attenuations and phase shifts. Such a channel is called *frequency selective*. In this case, multipath components separated by delay more than $1/W$ are resolvable and the channel impulse response can be written as [4]

$$h(t;\tau) = \sum_{p=1}^{P} h_p(t)\delta\left(\tau - \frac{p}{W}\right) \tag{4}$$

Since the multipath delay spread is T_m and the time resolution of multipaths is $1/W$, the number of paths L is given by $\lfloor T_m W \rfloor + 1$. In general, the time-varying tap coefficients $h_p(t)$ are modeled as mutually uncorrelated wide-sense stationary processes. The random time variations of the channel are generally modeled via a probability distribution on the channel coefficients $h_p(t)$. The most commonly used probability distributions are Rayleigh, Ricean, and the Nakagami-m [4].

The main purpose of the channel modeling is to characterize the channel in a tractable yet meaningful manner, to allow design and analysis of the communication algorithms. Note that *all* models are approximate representations of the actual channel, and thus development of practical systems requires both theoretical analysis and field testing.

In the sequel, we will consider only slowly fading channels, where $T_s \ll T_{\text{coh}}$, that is, multiple consecutive symbols or equivalently, a block of symbols undergo the same channel distortion. Hence, these channels are also referred as *block fading channels* [19–22]. As a result of slow time variation of the channel, the time dependency of the channel will be suppressed; that is, $h(t)$ will be denoted by h and $h(t;\tau)$ by $h(\tau)$.

4. CAPACITY OF MULTIPLE-ACCESS CHANNELS

Developed in the landmark paper by Shannon [23], information theory forms the mathematical foundation for source compression, communication over noisy channels and cryptography. Among other important contributions [23], the concept of *channel capacity* was developed. It was shown that a noisy channel can be characterized by its capacity, which is the maximum rate at which the information can be transmitted reliably over that channel. Information theoretic methods provide not only the ultimate achievable limits of a communication system but also valuable insights into the design of practical systems.

Typically, a capacity analysis starts by using a simple model of the physical phenomenon. The simplified model captures the basic elements of the problem, such as time-varying fading wireless channel, shared multiple access, and power-limited sources. Information-theoretic analysis then leads to limits on reliably achievable data rates and provides guidelines to achieve those limits. Although information-theoretic techniques are rarely practical, information-theory-inspired coding, modulation, power control and multiple-access methods have led to significant advances in practical systems. Furthermore, the analysis techniques allow performance evaluation of suboptimal but implementation-friendly techniques, thereby providing a useful benchmarking methodology.

In this section, we will provide a brief sampling of results pertaining to time-varying fading wireless channels; the reader is referred to Ref. 19 for a detailed review. Our aim is to highlight basic single and multiuser results for fading channels to motivate the algorithms discussed in the sequel. In Section 4.1, we will first introduce two notions of channel capacity, Shannon theoretic capacity [23] and outage capacity [24]. Capacity of a channel characterizes its performance limits using *any* practical transmitter–receiver pair and is a fundamental notion in evaluating efficacy of practical systems. Single-user fading channels will be analyzed using the two capacity notions, motivating the importance of diversity techniques (e.g., spacetime coding and beamforming) and power control. In Section 4.2, the multiuser extensions will be discussed to motivate the use of power-controlled CDMA-based multiple access.

All results in this section will be given for flat fading channels. The results can be easily extended to frequency-selective fading by partitioning the channel into frequency bins of width B_{coh}, and then treating each bin as a separate channel.

4.1. Capacity of Single User Fading Channels

A channel is deemed *noisy* if it introduces random perturbations in the transmitted signals. In Ref. 23, the capacity of a noisy channel was defined as the highest data rate at which reliable communication is possible across that channel. *Communication reliability* is defined as the probability that the receiver will decode the transmitted message correctly; higher reliability means lower errors in decoding messages and vice versa. An information rate is *achievable* if there exists at least one transmission scheme such that any preset level of communication reliability can be achieved. To achieve this (arbitrary level of) reliability, the transmitter can choose any codebook to map information message sequences to channel inputs. If the rate of transmission R is no more than the channel capacity C, then reliable communication is possible by using codebooks that jointly encode increasingly longer input messages. This notion of channel capacity is commonly referred as *Shannon theoretic capacity*.

Besides providing a characterization of the channel capacity for a broad class of channels, Shannon [23] also computed the capacity of the following additive white Gaussian noise (AWGN) channel,

$$z(t) = s(t) + v(t) \tag{5}$$

as

$$C = W \log_2\left(1 + \frac{\mathcal{P}_{\text{av}}}{\sigma^2}\right) \text{ bits per second (bps)} \tag{6}$$

Note that the AWGN channel in (5) can be considered as a special case of fading channel (3) with $h(t) \equiv 1$. In (6), W represents the channel bandwidth (in hertz), $\mathcal{P}_{\text{av}} = \mathbb{E}_s\{|s(t)|^2\}$ is the average transmitted power over time,[5] and σ^2 is the variance of the additive noise $\nu(t)$. The fundamental formula (6) clarifies the role of two important system parameters: the channel bandwidth W and signal to noise ratio (SNR), $\mathcal{P}_{\text{av}}/\sigma^2$. The capacity result (6) claims a surprising fact that even for very small amount of power or bandwidth, information can be sent at a nonzero rate with vanishingly few decoding errors. To achieve this reliable communication, the transmitter *encodes* multiple information bits together using a channel code. The encoded bits are then jointly decoded by the receiver to correct errors introduced by the channel (5).

The capacity analysis in Ref. 23 forms the basis for deriving capacity of fading channels (3), which we review next. With an average transmitted power constraint, $\mathbb{E}_s\{|s(t)|^2\} \le \mathcal{P}_{\text{av}}$, the Shannon theoretic capacity of fading channels, with perfect channel information at the receiver, is given by [25]

$$C_{sc}^{r} = W\mathbb{E}_{\gamma}\left\{\log_2\left(1 + \frac{\mathcal{P}_{\text{av}}\gamma(t)}{\sigma^2}\right)\right\} \tag{7}$$

where σ^2 is the variance of the additive i.i.d. Gaussian noise $\nu(t)$ in (3), and $\gamma(t) = |h(t)|^2$ is the received instantaneous power. The expectation in (7) is computed with respect to the probability distribution of the variable $\gamma(t)$. If, in addition to perfect channel information at the receiver, the transmitter has knowledge of the instantaneous channel realization, then the transmitter can adapt its transmission strategy based on the channel. The optimal strategy, in this case, turns out to be "water-filling" in time [26]. To water-fill in time, the transmitter waits for the good channel conditions to transmit and does not transmit during poor channel conditions. Thus, the optimal transmission policy is a constant rate Gaussian codebook (see Ref. 11 for details on Gaussian codebooks) transmitted using an instantaneous channel SNR-dependent power. The optimal transmission power is given by [26]

$$\mathcal{P}_{sc}(\gamma(t)) = \begin{cases} \mathcal{P}_{\text{av}}\left(\dfrac{1}{\gamma_{sc}} - \dfrac{1}{\gamma(t)}\right), & \gamma(t) \ge \gamma_{sc} \\ 0, & \gamma(t) < \gamma_{sc} \end{cases} \tag{8}$$

where the threshold γ_{sc} is found to satisfy the power constraint $\mathbb{E}_{\gamma,s}\{\mathcal{P}_{sc}(\gamma(t))|s(t)|^2\} \le \mathcal{P}_{\text{av}}$. The achievable capacity is then given by

$$C_{sc}^{rt} = W\mathbb{E}_{\gamma}\left\{\log_2\left(1 + \frac{\mathcal{P}_{sc}(\gamma(t))\gamma(t)}{\sigma^2}\right)\right\} \tag{9}$$

Note that allocated power in (8) is zero for poor channels whose SNR is less than $\gamma_{sc}(t)$ and increases monotonically as channels conditions improve. Adapting the transmission power based on channel conditions is known as *power control*. Channel state information at the transmitter leads to only modest gains for most fading distributions [26] with a single transmitter and receiver; thus C_{sc}^{rt} is only marginally greater than C_{sc}^{r}. But the gains of transmitter information increase dramatically with multiple transmit and receive antennas. Using the extensions of (7) and (9) to multiple antennas [25,27], a representative example is shown in Fig. 4. Thus, building adaptive power control policies is more useful for multiple antenna systems; see Ref. 28 for practical methods to achieve a significant portion of this capacity in a practical system. The gain due to channel state information at the transmitter can also be achieved by using imprecise channel information [28–30]. The large gains promised by multiple antenna diversity, with or without channel information at the transmitter, have sparked the rich field of space-time coding [5,31,32].

In slow-fading channels, achieving Shannon theoretic capacity requires coding over exceedingly long input blocks. The long codewords are required to average over different fading realizations, which then allow the use of assumed ergodicity[6] of the fading process to prove the capacity theorem. The large delays associated with Shannon theoretic capacity directly translate into impractical delays in delay sensitive applications like voice and video. Thus, with a delay constraint, the Shannon

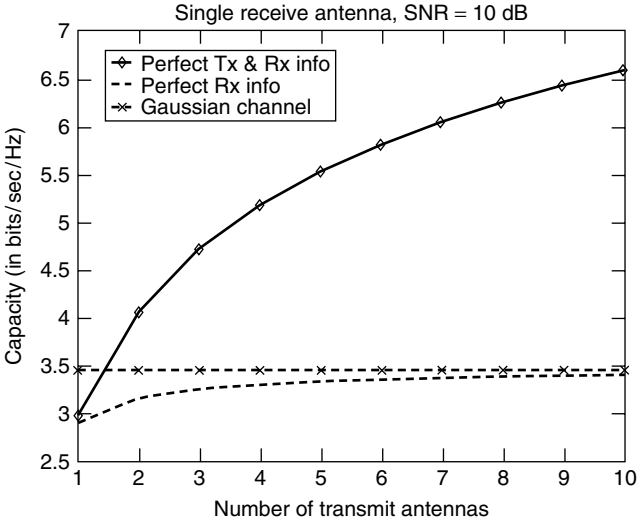


Figure 4. Capacity with multiple transmit antennas and single receive antenna, with different amount of channel state information at the transmitter.

[5] The expectation $\mathbb{E}_s\{|s(t)|^2\}$ represents an average computed over time (assuming that it exists) using the distribution of $s(t)$.

[6] A stochastic process $h(t)$ is called ergodic if its ensemble averages equal appropriate time averages [33]. The channel capacity theorem proved by Shannon [23] relied on the law of large numbers, where the time averages converge to their ensemble averages, which in turn motivated the idea of encoding increasingly long blocks of input messages. Ergodic channels are the most general channels with dependency across time for which the (strong) law of large numbers holds, thereby allowing a direct extension of capacity theorem [23] to ergodic channels. For a more general capacity theorem without any assumptions on channel structure, see Ref. 34.

theoretic capacity of slowly fading practical channels (more specifically, nonergodic channels) is zero [24]. In Ref. 24, the concept of capacity versus outage was introduced, which captures the effect of delay in slow fading channels. A block of transmitted data, which is assumed to undergo the same fading throughout, is in *outage* if the instantaneous capacity of the channel is less than the rate of transmission. The concept of outage provides a code-independent method (by using asymptotic approximations) to gauge the codeword error probability for practical codes. Assuming that the flat-fading channel h is constant for a block of transmitted data, the instantaneous capacity is given by[7] $W\log_2(1+\mathcal{P}_{\text{av}}\gamma(t)/\sigma^2)$. The outage probability, when only the receiver is aware of the channel state, is then given by

$$\Pi_{oc}^{r} = \text{Prob}\left(W\log_2\left(1+\frac{\mathcal{P}_{\text{av}}\gamma(t)}{\sigma^2}\right) < R\right) \tag{10}$$

where the probability is computed over the distribution of channel $h(t)$. Analogous to the preceding Shannon theoretic capacity analysis, the probability of outage can also be computed for different amount of channel state information at the transmitter. With perfect channel state information at the transmitter and receiver, the outage probability is given by

$$\Pi_{oc}^{rt} = \min_{\mathcal{P}_{oc}(\gamma(t))} \text{Prob}\left(W\log_2\left(1+\frac{\mathcal{P}_{oc}(\gamma(t))\gamma(t)}{\sigma^2}\right) < R\right) \tag{11}$$

The power allocation $\mathcal{P}_{oc}(\gamma(t))$ minimizing the outage is given by [27]

$$\mathcal{P}_{oc}(\gamma(t)) = \begin{cases} \dfrac{\sigma^2(2^{R/W}-1)}{\gamma(t)}, & \gamma(t) \geq \gamma_{oc} \\ 0, & \gamma(t) < \gamma_{oc} \end{cases} \tag{12}$$

The threshold γ_{oc} is chosen to meet the average power constraint, $\mathbb{E}_{\gamma,s}\{\mathcal{P}_{oc}(\gamma(t))|s(t)|^2\} \leq \mathcal{P}_{\text{av}}$. The *outage capacity*, which measures the total number of transmitted bits per unit time not suffering an outage, is given by

$$C_{oc}^{r} = (1-\Pi_{\text{out}}^{r})R$$
$$C_{oc}^{rt} = (1-\Pi_{\text{out}}^{rt})R$$

Because of the extra information at the transmitter, it immediately follows that $\Pi_{\text{out}}^{rt} < \Pi_{\text{out}}^{r}$ and hence $C_{oc}^{rt} > C_{oc}^{r}$. The gain in outage capacity due to transmitter information is much more substantial compared to Shannon capacity even for a single-antenna system [35]. Similar to the Shannon-capacity, outage capacity increases with the increasing number of transmit and receive antennas [25,36].

The differences in the objectives of achieving outage capacity versus achieving Shannon theoretic capacity can be better appreciated by the difference in the optimal power allocation schemes, $\mathcal{P}_{sc}(\gamma(t))$ and $\mathcal{P}_{oc}(\gamma(t))$. In the Shannon theoretic approach, the transmitter uses more power in the good channel states and less power during poor channel conditions. On the other hand, to minimize outage the transmitter employs *more* power as the channel gets *worse*, which is exactly opposite to the power allocation $\mathcal{P}_{sc}(\gamma(t))$. The difference in power allocation strategies, $\mathcal{P}_{sc}(\gamma(t))$ and $\mathcal{P}_{oc}(\gamma(t))$ can be attributed to optimization goals: Shannon theoretic capacity maximizes long-term throughput and hence it is not delay-constrained, and outage capacity maximizes short-term throughput with delay constraints.

Irrespective of the capacity notion, the main lesson learned from information-theoretic analysis is that diversity and channel information at the transmitter can potentially lead to large gains in fading channels. The gains promised by above information-theoretic results have motivated commonly used methods of space-time coding and power control to combat fading. Readers are referred to the literature [21,25,26,36–38] for detailed results on capacity of single user flat-fading channels. In the next section, we will briefly discuss the results for multiple access channels and their impact on the choice of multiple access protocols.

4.2. Multiple User Fading Channels

The primary question of interest in a multiuser analysis is the multiaccess protocol to efficiently share the spectral resources among several power-limited users. An accurate capacity analysis of a complete cellular system is generally intractable. Hence, the information-theoretic analysis relies on a series of simplifying assumptions to understand the dominant features of the problem. Our main emphasis will be on uplink communication in a single cell, where multiple users simultaneously communicate with a single receiver, the base station.

The sampled received baseband signal at the base station is the linear superposition of K user signals in additive white Gaussian noise, given by

$$y(t) = \sum_{i=1}^{K} h_i(t)s_i(t) + \nu(t) \tag{13}$$

The Gaussian noise $\nu(t)$ is assumed to be zero mean with variance σ^2. The channels for all users $h_i(t)$ are assumed to vary independently of each other and from one coherence interval to another. The fading processes for all users are assumed to be jointly stationary and ergodic. Furthermore, each user is subjected to an average power constraint, $\mathbb{E}_{s_i}\{|s_i(t)|^2\} \leq \mathcal{P}_i$.

Equivalent to the capacity of channel in the single-user case, a *capacity region* specifying all the rates that can be simultaneously and reliably achieved are characterized. Thus, the capacity region for K users is a set of rates defined as

$$\mathcal{R} = \{\underline{R} = (R_1, R_2, \ldots, R_K) : \text{ rates } R_i \text{ can be reliably achieved simultaneously}\} \tag{14}$$

When the base-station receiver is aware of all the fading realizations of all the users, $\{h_i(t)\}$, then the rate region

[7] Assuming that the transmitter is unaware of the instantaneous channel state and receiver has the perfect knowledge of $h(t)$ [25].

is described by the following set of inequalities (in the single-user case, there is only one inequality, $R \le C$)

$$\sum_{i\in\mathcal{B}} R_i \le \mathbb{E}_{\underline{\gamma(t)}} \log_2\left(1+\frac{\sum_{i\in\mathcal{B}}\gamma_i(t)\mathcal{P}_{\text{av}}}{\sigma^2}\right) \tag{15}$$

where it is assumed that each user has the same average power limit $\mathcal{P}_i = \mathcal{P}_{\text{av}}$. In (15), $\mathcal{B}$ represents a subset of $\{1, 2, \ldots, K\}$, $\gamma_i(t) = |h_i(t)|^2$ is the received power, and $\gamma(t) = [y_1(t)y_2(t)\ldots y_k(t)]$. The expectation of $\mathbb{E}_{\underline{\gamma(t)}}$ is over all the fading states $\{\gamma_i(t)\}_{i\in\mathcal{B}}$. A quantity of interest is the *normalized sum rate*, which is the maximum achievable equal rate per user and is obtained by taking $\mathcal{B}$ to be the whole set to yield [39]

$$R_{\text{sum}} = \frac{1}{K}\sum_{i=1}^{K} R_i = \mathbb{E}_{\underline{\gamma(t)}}\frac{1}{K}\log_2\left(1+\frac{\mathcal{P}_{\text{av}}\sum_{i=1}^{K}\gamma_i(t)}{\sigma^2}\right) \tag{16}$$

$$\underset{K\to\infty}{\to} \frac{1}{K}\log_2\left(1+\frac{K\mathcal{P}_{\text{av}}}{\sigma^2}\right) \tag{17}$$

The asymptotic result (17) shows an interesting phenomenon, that as the number of users increases, the effect of fading is completely mitigated because of the averaging effect of multiple users. The averaging effect due to increasing users is analogous to time or frequency [40] or spatial [25] averaging in single-user channels. Shamai and Wyner [39], using (16), showed that a nonorthogonal multiple-access scheme has a higher normalized sum rate R_{sum} than orthogonal schemes such as time (frequency) division multiple access.[8] By requiring orthogonality of users, an orthogonal multiple access scheme adds additional constraints on user transmission, which leads to a performance loss compared to optimal nonorthogonal method. Nonorthogonal CDMA is an example of the nonorthogonal multiple-access scheme. Spread signals, like CDMA signals, occupy more bandwidth than needed and were first conceived to provide robustness against intentional jamming [41]. The capacity–outage analysis also shows the superiority of CDMA schemes over orthogonal access methods [42].

A cellular multicell model [43] was introduced to study the effect of multiple cells. The model extends (13) to include intercell interference from users in neighboring cells. The cellular model [43] was extended to fading channels [39,44]. There again, it was concluded that CDMA, like wideband methods achieve optimal normalized sum rates even in the presence of multicell interference, for several important practical receiver structures. Even though the spread-spectrum signals occupy more bandwidth than needed for each signal, multiuser spread spectrum systems are spectrally efficient [13,14]. Motivated by the success of the second generation CDMA standard, IS-95, currently all third generation wireless systems (CDMA2000 and W-CDMA) use some form of spread spectrum technique. In addition to information theoretic superiority, CDMA-based multiple access provides other practical advantages [45]. First, CDMA signals allow finer *diversity combining* due to larger signal bandwidth, thereby providing robustness to multipath fading. In other words, combined with an interleaver, spread-spectrum signals naturally exploit both frequency and time diversity. Frequency diversity is not available in bandwidth-efficient TDMA systems. Moreover, CDMA allows a *frequency reuse* of one in contrast to TDMA/FDMA, which requires a higher reuse factor. A lower reuse factor immediately implies higher system capacity; a reuse factor of one also simplifies frequency planning. Finally, CDMA naturally exploits the *traffic activity factor*, the percentage of time during a two-way communication each channel is actually used. Most of the information theoretic analysis completely ignores the data burstiness, a property which is central to higher resource utilization in wired networking [46]; see Refs. 47 and 48 for insightful reviews.

The CDMA based systems allow communication without the need for a universal clock or equivalently synchronism among different users. The need for synchronism in TDMA requires the use of time guard bands between time slots and hence wastes resources. Finally, in long-code DS-CDMA systems, like the one used in IS-95 standard[9] assigning channels to users is straightforward because each user is given a unique fixed spreading code. In TDMA, time slots are granted adaptively as users hand off from one cell to another, thereby complicating resource management and requiring additional protocol overhead. Also, long-code CDMA leads to the same average performance for all users, and thus a fair resource allocation among users.

Although the area of multiuser information theory is rich and well studied, we maintain that many fundamental results are yet to be published. For instance, connections with queuing theory [47–49], which is the mathematical basis for networking, are far from well understood, but with the rise of Internet, it is more urgent than ever to unify the areas of data networking and wireless communications. Furthermore, with the growth of wireless services beyond voice communication, and advent of newer modes of communication like ad hoc networking,[10] current information-theoretic results should be considered as the beginnings of our understanding on the subject of multiuser communications.

5. TYPICAL ARCHITECTURE OF WIRELESS TRANSCEIVER

Most wireless systems transmit signals of finite bandwidth using a high-frequency carrier.[11] This immediately leads to the wireless transceiver with three major components: (1)

[8] In time (frequency)-division multiple access, each user transmits in its allocated time (frequency) slot such that no two users share a time (frequency) slot. Thus, the transmission of one user is orthogonal in time (frequency) to any other user.

[9] In long-code CDMA systems, unlike short repeating-code CDMA systems, each transmitted bit is encoded with a different spreading code.

[10] In ad hoc networking, mobile nodes can communicate with each other without the need for any infrastructure as in cellular systems; IEEE 802.11 and Bluetooth are examples of ad hoc networking.

[11] Carrierless systems include impulse radio [50].

an RF front end that performs the frequency conversion from passband to baseband and vice versa, (2) digital-to-analog converter (D/A) and analog to digital (A/D) converter, and (3) a baseband processing unit. In this section, we will discuss the signal processing algorithms used in the digital baseband unit. Wherever applicable, we will highlight the differences between the baseband unit at the mobile receiver and that at the base station.

We briefly note that the hardware receiver design for CDMA systems is generally more challenging than its TDMA counterparts. The design of A/D, D/A converters, and digital baseband processors require special effort. Higher chipping rates in CDMA systems require faster sampling and hence lead to higher computational throughout requirements and increased circuit power dissipation compared to their TDMA counterparts. Fortunately, advances in low-power high-speed complementary metal oxide semiconductor (CMOS) circuits have allowed implementation of sophisticated digital signal processing algorithms, and high-speed converters.

5.1. Transmitter

A simplified transmitter for DS-CDMA system is shown in Fig. 5. The data obtained from the higher layers is passed through a channel encoder, spread-spectrum modulator, digital to analog converter and finally through an RF unit.

5.1.1. Channel Encoding. The source data bits are first encoded using a forward error correction (FEC) code. A FEC code systematically adds redundant bits to the source bits, which are used by the receiver to correct errors in the received signal. Error correction coding is essential to achieve low bit error rates at the receiver and has a strong information theoretic foundation [23]. Following Shannon's work in 1948 [23], error control coding has seen tremendous growth since 1950; the readers are referred to the literature [51–54] for recent reviews on state of the art. Several excellent texts [55–58] on channel coding theory are available, hence we will keep our discussion in this section elementary.

The choice of code primarily depends on desired performance level, the specific channel under consideration and the complexity of the resulting receiver. The desired level of performance is based on the type of services to be provided. For instance, loss tolerant services such as speech can work with high packet loss probability, while data/email/fax requires a much higher error protection, thereby requiring FEC codes with different amount of error protection capabilities.[12] The complexity of decoding the received packets to correct errors is a major concern in the design of power-limited mobile handsets. Typically, stronger FEC codes are computationally harder to decode and, hence require more battery power for the baseband units; see Ref. 59 for a discussion.

The communication channel is a major factor in selection of FEC codes. For example, code design is different for slow- and fast-fading channels. To illustrate the concept of coding, our discussion will be limited to convolutional codes that are used in both telephone line modems, and both second and third generation digital wireless cellular standards. Further, we will highlight the interest in space-time coding by dividing this section into two parts: single-antenna systems and multiple-antennas systems. Our discussion on single-antenna systems will give a quick introduction to convolutional codes with a review of the most recent coding results for slow and fast fading channels. In the multiple antenna discussion, diversity techniques will be central to our discussion, with an emphasis on spatial and time diversity for wireless systems.

5.1.1.1. Single-Antenna Systems. The choice of convolutional codes is motivated by their simple optimal decoding structure, systematic construction of strong codes for large block lengths, and lower decoding delay compared to block codes. A convolutional code is generated by passing the information sequence through a linear finite-state shift register. In general, the shift register consists of S B-bit stages and m linear algebraic function generators; see Fig. 6 [4]. The input data to the encoder, assumed to be binary, are shifted into and along the shift register B bits at a time. The number of output bits for each B input bits is m bits. Consequently, the code rate is defined as $R_c = B/m$. The parameter S is called the *constraint length* of the convolutional code.

To understand the encoding procedure, consider the convolutional encoder for $S = 3$, $B = 1$, and $m = 3$ shown in Fig. 7 [4]. All the shift registers are assumed to be in

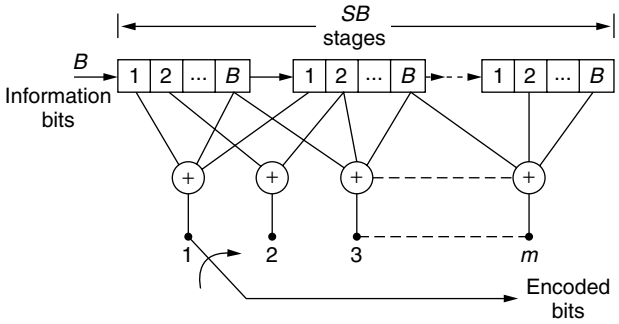


Figure 6. Convolutional encoder.

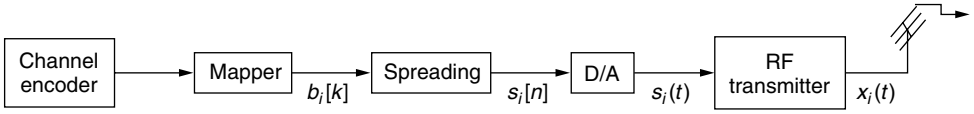


Figure 5. DS-CDMA handset transmitter components.

[12] Some of the networking layers use checksums for error detection and perform error correction by requesting retransmission of packets.

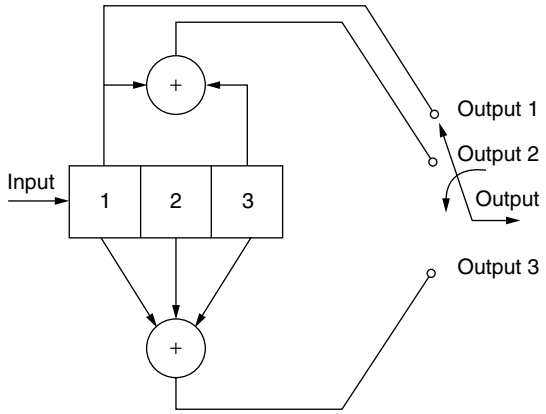


Figure 7. Convolutional encoder for a (3,1) code.

zero state initially. If the first input bit is a 1, the resulting output sequence of 3 bits is $[b[1]\ b[2]\ b[3]] = [1\ 1\ 1]$. Now, if the second input bit is a 0, the next three output bits are $[b[4]\ b[5]\ b[6]] = [0\ 0\ 1]$ (or else the output bits are [110] if the input bit is 1). If the third bit is a 1, the output is $[b[7]\ b[8]\ b[9]] = [1\ 0\ 0]$. The operation of a nonrecursive (Fig. 7) convolutional code is similar to that of a finite-impulse response (FIR) filter with all the operations done over a finite field; in Fig. 7, the finite field consists of only two elements $\{0, 1\}$ with binary addition. The convolutional code has one input and several outputs, equivalent to a single-input multiple-output FIR linear system. The equivalent of the impulse response of the filter is the *generator polynomial*, which succinctly describes the relation between output and shift register states for a convolutional code. For the example in Fig. 7, the generator polynomials are

$$\text{Output } 1 \rightarrow \mathbf{g}_1 = [1\ 0\ 0]$$

$$\text{Output } 2 \rightarrow \mathbf{g}_2 = [1\ 0\ 1]$$

$$\text{Output } 3 \rightarrow \mathbf{g}_3 = [1\ 1\ 1]$$

The generator polynomials of a convolutional code characterize its performance via different metrics, notably minimum distance and distance spectrum [60]. To design any code requires an appropriate metric space, which depends on the channel under consideration. For slowly block fading channels, the Euclidean distance between the codewords is the natural metric [60], while for fast fading channels, Hamming distance is the appropriate metric [61]; see discussion below for further discussion on diversity techniques.

Addition of redundant bits for improving the error probability leads to bandwidth expansion of the transmitted signal by an amount equal to the reciprocal of the code rate. For bandwidth constrained channels, it is desirable to achieve a coding gain with minimal bandwidth expansion. To avoid bandwidth expansion due to channel coding, the number of signal points over the corresponding uncoded system can be increased to compensate for the redundancy introduced by the code. For instance, if we intend to improve the performance of an uncoded system using BPSK modulation, a rate $\frac{1}{2}$ code would require doubling the number of signal points to quadrature phase shift keying (QPSK) modulation. However, increasing the number of signals leads to higher probability of error for the same average power. Thus, for the resultant bandwidth efficient scheme to provide gains over the uncoded system, it must be able to overcome the penalty due to increased size of the signal set.

If the modulation (mapping of the bits to channel signals) is treated as an operation independent of channel encoding, very strong convolutional codes are required to offset the signal set expansion loss and provide significant gains over the uncoded system [4]. On the other hand, if the modulation is treated as an integral part of channel encoding, and designed in unison with code to maximize the Euclidean distance between pairs of coded signals, the loss due to signal set expansion is easily overcome. The method of *mapping by set partitioning* [62] provides an effective method for mapping the coded bits into signal points such that the minimum Euclidean distance is maximized. When convolutional codes are used in conjunction with signal set partitioning, the resulting method is known as *trellis-coded modulation* (TCM). TCM is a widely used bandwidth efficient coding scheme with a rich associated literature; see Ref. 63 for a comprehensive in-depth review.

The fundamental channel coding theorem by Shannon [23] proved the existence of good codes, which can achieve arbitrarily small probability of error, as long as the transmission rate is lower than the channel capacity. The proof in Ref. 23 required creating codes that had continually increasing block sizes to achieve channel capacity. Another key component of the proof in Ref. 23 was the choice of codebooks, they were chosen at random. Random codes with large block sizes have no apparent structure to implement a physically tractable decoder. Proven optimality of random codes coupled with the inability to find good structured codes led to a common belief that the structured deterministic codes had a lower capacity than the channel capacity, often called the "practical capacity" [64,65]. The discovery of *turbo codes* [17] and the rediscovery of *low-density parity-check* (LDPC) codes [66] appears to have banished the abovementioned "practical capacity" myth. Both Turbo and LDPC codes have been shown to operate below the "practical capacity," within a tenth of a decibel of the Shannon capacity. Turbo codes have also been proposed for the third-generation wireless standards. The main ingredients of a turbo code are constituent codes (block or convolutional code) and a long interleaver. The long interleaver serves two purposes: lends codewords a "randomlike" structure, and leads to long codes that are easily and efficiently decoded using a (suboptimal yet effective) iterative decoding algorithm. Several extensions of turbo codes are areas of active research, notably, bandwidth-efficient Turbo codes [67,68], deterministic interleaver design [69] and spacetime Turbo codes [70].

We close the discussion on codes for slow fading Gaussian channels, by highlighting that none of the current codes come close to the lower bounds on the performance of codes [71]. Current codes require large block lengths to achieve small probability of decoded

message errors, but relatively short block lengths suffice to achieve the same level of performance for "good" codes [71]. Thus, the field of code design, although more than fifty years old, has still significant room to develop.

5.1.1.2. Multiple-Antenna Systems. The random time variations in the received signal provide diversity, which can be exploited for improved error performance. Typical forms of diversity include time, frequency, and spatial diversity. In Section 4.1, it was noted that diversity is important to improve the outage performance or achievable rates in fading channels. Although only spatial diversity using multiple transmit and receive antennas was studied in Section 4.1, similar benefits are also obtained by using time or frequency diversity or a combination of them. In time and frequency diversity, channel variations in time and across frequency are used to increase reliability of the received signal. In spatial diversity, multiple transmit and/or receive antennas exploit the random spatial time variations.

The codes designed for Gaussian channels can be used for slowly fading channels if an accurate channel estimator is available and all symbols of a codeword undergo the same channel fading. In the presence of medium to fast fading, where the coherence interval is shorter than a codeword, Hamming distance between the codewords should be maximized [61]. If channel variations are slower than a codeword, an interleaver is commonly used to induce time diversity. For interleaver-based schemes to be effective, the interleaver depths should be larger than the coherence interval; this implies that it is useful for fast-fading channels or for communications where large delay can be tolerated. For low-delay application, the interleaver-induced time diversity is not possible. In addition, if the channel is flat-fading (true for narrowband communications), then frequency diversity cannot be used, either. Irrespective of the availability of time and frequency diversity, the spatial diversity via multiple antennas is a promising method to achieve higher data rates.

Receiver diversity using multiple receive antennas is a well-understood concept [6] and often used in practice [72]. In contrast, using multiple antennas at the transmitter has gained attention only relatively recently due to discovery of space-time codes [5,31], motivated by encouraging capacity results [25,73]. Space-time coding exploits multiple independent channels between different transmit–receive antenna pairs in addition to time diversity (possibly interleaver induced). Later work [5] extended well-founded coding principles to spatial diversity channels, thereby simultaneously achieving coding gain and the highest possible spatial diversity. The space-time codes proposed there [5] have become a performance benchmark for all subsequent research in space-time coding [74–80]. The concept of transmitter diversity can be appreciated using the following elegant *Alamouti scheme* [81] for two transmit antennas.

In a given symbol period, two symbols are simultaneously transmitted from the two antennas. Denote the signal transmitted from antenna 1 as s_1 and from antenna 2 as s_2 (see Fig. 8). During the next symbol period, signal $-s_2^*$ is transmitted from antenna 1, and s_1^* is transmitted from antenna 2. Note that the encoding of symbols is done in both space and time. As is evident from Fig. 8, the received signal in any symbol interval is a linear combination of the signals transmitted from the two antennas. Thus, the spacetime channel is an interference channel. An analogous scenario exploiting frequency diversity would use nonorthogonal carrier frequency to send two symbols in each symbol period. The Alamouti scheme sends orthogonal signals over two time instants from the two antennas, where vector $[s_1 - s_2^*]$ transmitted from antenna 1 over two time symbols is orthogonal to the vector $[s_2 \;\; s_1^*]$ transmitted from antenna 2. If the channel stays constant over two consecutive symbol periods, then the orthogonality is maintained at the receiver. Since each symbol s_1 and s_2 is transmitted from both the antennas, they travel to the receiver from two different channels, which provides the desired diversity order of two. The orthogonality of the time signals helps resolution of the two symbols at the receiver without affecting the diversity order.

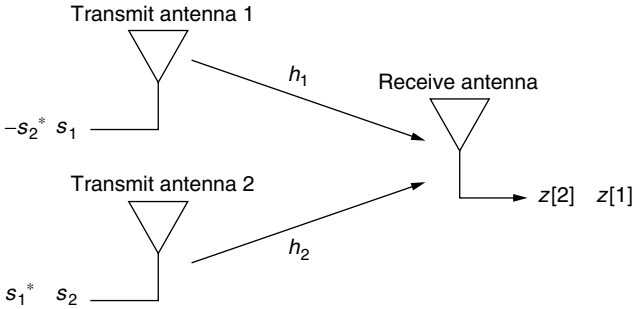


Figure 8. Alamouti encoder for two transmit and one receive antenna.

The Alamouti scheme can be extended to more than two transmit antennas using the theory of orthogonal designs [74]. The Alamouti scheme is a rate 1 code and thus requires no bandwidth expansion. But it provides a diversity order of two, which is twice that of any rate 1 single-antenna system. The Alamouti scheme has a very simple optimal receiver structure, thereby making it a prime candidate for practical implementations. In addition to its simplicity, the Alamouti scheme-based systems do not lose in their asymptotic performance. It has been shown [79] that orthogonal transmit diversity schemes are capacity-achieving, and this provided a motivation for the concatenated space-time coding methods [79,80]. The concatenated space-time codes decouple the spatial and temporal diversity to simplify the space-time code design.

All third-generation systems have adopted some form of transmit and receive diversity. Multiple antennas at the base station are relatively easier to implement in comparison to multiple antennas at the mobile handset, due to size limitation. Two cross-polarized antennas have been proposed and tested for mobile handsets [82].

5.1.2. Spreading and Modulation. The binary output of the error control encoder is mapped to either ± 1 to obtain the sequence $b_i[k]$, which is multiplied by a spreading sequence, $c_i[n] \in \{-1, 1\}$, of length N; the spreading operation is shown in Fig. 9. After spreading the

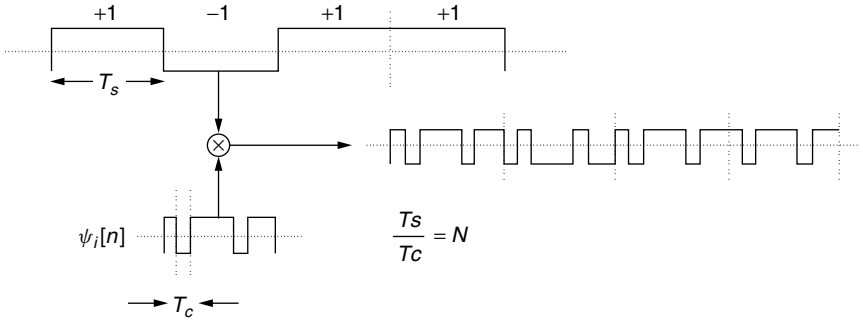


Figure 9. The spreading operation.

signal, the signal is passed through a digital pulseshaping filter, $\phi[n]$, which is typically a square-root raised-cosine filter [4]. The pulse shaping filter is chosen to limit the bandwidth of the transmitted signal to the available spectrum, while minimizing the intersymbol interference (ISI) caused by the filter. The digital signal for user i after pulse shaping can be written as

$$s_i[n] = \sum_{k=1}^{G} b_i[k]\psi_i[n - kNL] \tag{18}$$

where L is the number of samples per chip and $\psi_i[n] = \phi[n] * c_i[n]$, where $*$ represents linear convolution, and G is the number of bits in the packet.

After converting the digital signal to analog using a D/A converter, the RF upconverter shifts the baseband analog signal to the carrier frequency f_c. The upconverted signal is amplified by a power amplifier and transmitted via an antenna. The transmitted passband signal assumes the form

$$\begin{aligned} x_i(t) &= \sqrt{\mathcal{P}_i} e^{-j\omega_c t} \sum_{k=1}^{G} b_i[k]\psi_i(t - kT_s) \\ &= e^{-j\omega_c t} s_i(t) \end{aligned} \tag{19}$$

where T_s is the symbol period and $\mathcal{P}_i$ is the transmitted power. The bits $b_i[k]$ are the output of a suitable channel encoder discussed in Section 5.1.1. Since CDMA signals at the base station typically have large peak to average power ratios, the operating point of the power amplifier is kept low to avoid amplifier nonlinearities. The amplifier nonlinearities are avoided for several important reasons: (1) RF amplifier efficiency is lower in nonlinear region, which increases the power loss and hence total power consumed by the transmitter; (2) the nonlinearity introduces higher spectral components, which can cause increased interference in the neighboring frequency bands; and (3) the algorithm design for resulting nonlinear systems becomes intractable.

As discussed in Section 4.1, multiple antennas at the transmitter and receiver can lead to large gains in fading wireless channels [21,25,37]. If multiple transmit antennas are used, the vector transmitted passband signal is given by

$$\mathbf{x}_i(t) = \sqrt{\frac{\mathcal{P}_i}{M}} e^{-j\omega_c t} \sum_{k=1}^{G} \mathbf{b}_i[k]\psi_i(t - kT) \tag{20}$$

where M is the number of transmit antennas. The $M \times 1$ vectors, $\mathbf{x}_i(t)$ and $\mathbf{b}_i[k]$, represent the transmitted vector signal and spacetime-coded signal, respectively. In (20), we have assumed that the transmitter has no knowledge of the channel and hence uses the same average power on each transmitter. If the transmitter "knows" the channel, then the power across different antennas can be adapted to achieve an improved performance [25,83].

5.2. Base-Station Receiver

In cellular systems, the time and spectral resources are divided into different logical *channels*. The generic logical channels are broadcast, control, random-access, paging, shared, and dedicated channels [84,85]. All logical channels are physically similar and the distinction is solely made based on the purpose served by each channel. In the sequel, we will consider only the dedicated and shared channels, since they carry most of the user data and hence impose the biggest computational bottleneck. Implementation details of other channels can be found elsewhere [84,85].

As noted in Section 2, the unknown time-varying multipath is one of the biggest challenges in the design of wireless systems. Optimal transmission schemes that do not require knowledge of the wireless channel at the receiver can be designed using information theoretic tools (see Ref. 86 and the references therein), but are seldom employed. The primary reason for not using optimal strategies is their high computational complexity, and large latency of the resulting communication method. Hence, suboptimal and computationally efficient solutions are generally employed. The receiver estimates the unknown channel, and then uses the channel estimate to decode the data using a channel decoder.

A simplified illustration of the baseband receiver is shown in Fig. 10. The key components of the receiver are multiuser channel estimation, multiuser detection, and single-user channel decoding. Most systems also provide feedback from the receiver for power control and automatic repeat request (ARQ) to improve system reliability. The choice of algorithms used in each of the blocks is determined by their computational complexity, desired performance level, and the available side information. Mobile units are power- and complexity-constrained, and have little or no knowledge of the multiple access interference. On the other hand, the base stations are equipped with higher processing power and detailed

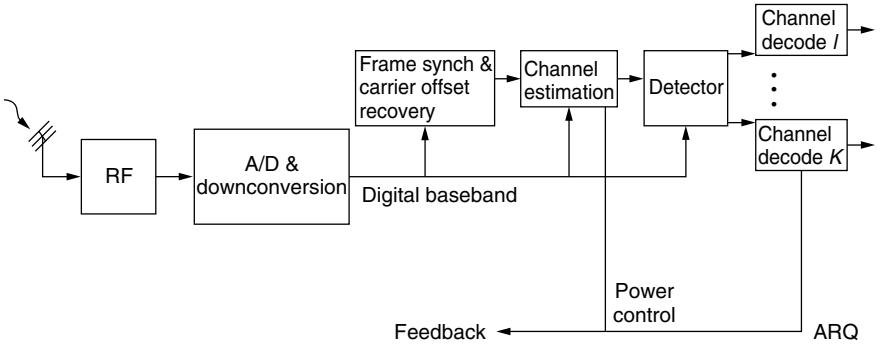


Figure 10. Base-station receiver structure.

information about all in-cell users, thereby allowing more sophisticated processing at the base stations. Our discussion will focus on base-station algorithms in the following section, with only bibliographic references to relevant counterparts for the mobile handset.

5.2.1. Received Signal. For each active user in a cell, the received signal at the base station consists of several unknown time-varying parameters. These parameters include propagation delay, amplitude, delay and number of paths, and residual carrier offset. The time variation in propagation delay is caused as users move closer or away from the base-station. The mobility of the users or the surrounding environment also causes time variation in the multipath environment. Finally, drift in the local oscillator frequencies of the transmitter and receiver leads to a residual carrier offset at the baseband.

Using the model (4) for the multipath channel impulse response and assuming that the channel coefficients for the ith user $h_{p,i}$ are constant over the observation interval, the received signal for a transmitted signal $x_i(t)$ without additive white noise is given by

$$\begin{aligned} z_i(t) &= \sum_{p=1}^{P} h_{p,i} x_i \left(t - \tau_i - \frac{p}{W}\right) \\ &= \sqrt{\mathcal{P}_i} e^{-j\omega_c t} \sum_{p=1}^{P} \underbrace{h_{p,i} e^{j\omega_c(\tau_i + p/W)}}_{a_{p,i}} \\ &\quad \times \sum_{k=1}^{G} b_i[k]\psi_i \left(t - kT - \tau_i - \frac{p}{W}\right) \\ &= \sqrt{\mathcal{P}_i} e^{-j\omega_c t} \sum_{k=1}^{G}\sum_{p=1}^{P} b_i[k] a_{p,i} \psi_i \left(t - kT - \tau_i - \frac{p}{W}\right) \end{aligned} \tag{21}$$

where τ_i is the propagation delay of the received signal. If the number of paths $P = 1$, then it is a flat-fading channel else a frequency selective channel. The received signal is amplified and downconverted to baseband. In practice, there is a small difference in the frequencies of the local oscillators at the transmitter and the receiver. The received baseband signal after downconversion (without additive noise) is given by

$$z_i(t) = \sqrt{\mathcal{P}_i} e^{-j\Delta\omega_i t} \sum_{k=1}^{G}\sum_{p=1}^{P} b_i[k] a_{p,i} \psi_i \left(t - kT - \tau_i - \frac{p}{W}\right) \tag{22}$$

where $\Delta\omega_i$ represents the residual carrier frequency offset. Assuming that the carrier offset $\Delta\omega_i$ is negligible or is corrected using a multiuser equivalent of digital phase-locked loop [4,87], the sampled baseband (without noise) with L samples per chip can thus be written as

$$z_i[n] = \sum_{k=1}^{G}\sum_{p=1}^{P} b_i[k] a_{p,i} \psi_i[n - kNL - \tau_i - p] \tag{23}$$

In general, the receiver components introduce thermal noise, which is generally modeled as additive noise. For K simultaneously active users, the received baseband signal in the presence of thermal noise at the base station is

$$z[n] = \sum_{i=1}^{K} z_i[n] + \nu[n] \tag{24}$$

The additive component $\nu[n]$ in (24) is generally modeled as white Gaussian noise. The received signal model in Eqs. (13) and (24) are similar; both consider a sum of all user signals in additive noise. The main difference is the assumption on the fading statistics; a flat-fading model is assumed in (13) compared to a multipath model in (24).

In the sequel, we will focus on estimating the unknown channel coefficients and subsequent detection of the data bits, $b_i[k]$ for all users $i = 1, \ldots, K$. The development of multiuser channel estimation and data detection is greatly simplified by using linear algebraic methods. We will write the received signal (24) using matrix–vector notation in two different forms. The first form will be used in multiuser channel estimation methods, and the second in multiuser detection.

5.2.1.1. Channel as Unknown. For simplicity, we will assume that all τ_i are multiple of sampling instants, $\tau_i = l_i$; for the general case, the reader is referred to Ref. 88. Let $u_i[n] = \sum_{k=1}^{G} b_i[k]\psi(n - kNL)$. Then the received signal $z_i[n]$

can be rewritten in matrix–vector notation [89] as

$$\mathbf{z}_i = \begin{bmatrix} u_i[1] & 0 & 0 & \cdots & 0 \\ u_i[2] & u_i[1] & 0 & & 0 \\ u_i[3] & u_i[2] & u_i[1] & & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \vdots & u_i[GLN + l_\phi] \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ a_{1,i} \\ \vdots \\ a_{P,i} \end{bmatrix}$$

$$= \mathbf{U}_i \mathbf{a}_i, \tag{25}$$

where there are l_i leading zeros in the channel vector $\mathbf{a}_i$ to account for the propagation delay, and l_ϕ is the length of the pulse ϕ (measured in number of samples). The total received signal can thus be written as

$$\mathbf{z} = [\,\mathbf{U}_1 \quad \mathbf{U}_2 \quad \cdots \quad \mathbf{U}_K\,] \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_K \end{bmatrix} + \boldsymbol{\nu}$$

$$= \mathbf{U}\mathbf{a} + \boldsymbol{\nu} \tag{26}$$

where we recall that N is the spreading gain, L is the number of samples per chip, and G is the number of bits in the packet. The signal model described above will be used to derive channel estimation algorithms in Section 5.2.2.

5.2.1.2. Data as unknown. Define $q_i[n] = \sum_{p=1}^{P} a_{p,i}\,\psi[n - kNP - l_i - p]$; $q_i[n]$ can be understood as the *effective* spreading waveform for the ith user. The waveform $q_i[n]$ is generally longer than one symbol period and hence causes interference between the consecutive symbols. To highlight the presence of intersymbol interference (ISI), we will write the received signal $z_i[n]$ for every symbol duration. For simplicity, we will assume that the length of $q_i[n], l_q$, is less than two symbol durations, namely, $l_q < 2NL$. Then the received signal $z_i[n]$ can be written as

$$\mathbf{z}_i[k] = \begin{bmatrix} 0 & q_i[1] \\ 0 & q_i[2] \\ \vdots & \vdots \\ q_i[NL+1] & q_i[2NL - l_q + 1] \\ \vdots & \vdots \\ q_i[l_q] & q_i[NL] \end{bmatrix} \begin{bmatrix} b_i[k-1] \\ b_i[k] \end{bmatrix}$$

$$= \mathbf{Q}_i \mathbf{b}_i \tag{27}$$

The total received signal can be written as

$$\mathbf{z}[k] = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 & \cdots & \mathbf{Q}_K \end{bmatrix} \begin{bmatrix} \mathbf{b}_1[k] \\ \mathbf{b}_2[k] \\ \vdots \\ \mathbf{b}_K[k] \end{bmatrix} + \boldsymbol{\nu} \tag{28}$$

$$= \mathbf{Q}\mathbf{b}[k] + \boldsymbol{\nu} \tag{29}$$

The received signal model in (29) clearly demonstrates the challenges in multiuser detection. Not only does the receiver have to cancel the multiple access interference, but also the ISI for each user introduced by the multipath channel. The ISI acts to increase the effective multiple access interference experienced by each bit. The multiuser detection methods aim to jointly make all bits decisions $\mathbf{b}[k]$.

In the following section, we will discuss channel estimation, multiuser detection and channel decoding algorithms for DS-CDMA systems.

5.2.2. Multiuser Channel Estimation. Most channel estimation can be divided into two broad classes: training based and blind methods. In each class, a further subdivision[13] is made on the basis of assumptions made regarding the multiple access interference: single-user channel estimation in the presence of multiple access interference or jointly estimating channels for all the users.

Most wireless systems add known symbols periodically to the data packets. The known data symbols are known as *training symbols* and facilitate coarse synchronization, channel estimation and carrier offset recovery. Training-based methods simplify estimation of unknown baseband parameters at the cost of throughput loss; symbols used for training could potentially be used to send more information bits. The amount of training depends on the number of simultaneous users, the number of transmit antennas [28], and the desired reliability of channel estimates. Given the training symbols and assuming perfect carrier offset recovery, multiuser channel estimation can be cast as a linear estimation problem [91], and admits a closed-form solution. The work in Ref. 91 also discusses extensions to multiple antennas.

A class of blind channel estimation procedures, collectively known as *constant modulus algorithms* (CMA), were first proposed [92,93] using the constant amplitude property of some of the communication signals such as BPSK. The CMA algorithms use a nonlinear (nonconvex) cost function to find the channel estimate, and hence can converge to poor estimates. An alternate procedure of blind estimation was proposed [94,95], which used the cyclostationarity of the communication signals. Motivated by another method [94], a single-user blind channel identification method, using only second-order statistics, was proposed [96]. The blind channel equalization exploits only the second (or higher)-order statistics without requiring periodic training symbols, with an assumption that the data symbols are independent and identically distributed. The assumption of i.i.d. data is rarely correct because of channel coding used in almost all systems. Hence, the results based on blind channel estimation should be interpreted with caution. Nonetheless, there is value in exploring blind channel identification methods. Blind estimation can improve the estimates based on training or completely avoid the use of training symbols; the reader is referred elsewhere [97,98] for results on single-user systems.

[13] Another possible subdivision can be based on linear and nonlinear algorithms. An example of feedback-based nonlinear algorithm is the decision-feedback-based equalization [90].

Single-user channel estimation in the presence of unknown multiple access interference has been addressed [99]. An approximate maximum-likelihood channel estimation for multiple users entering a system has been presented [100]; the estimate-maximize algorithm [101] and the alternating projection algorithm [102] in conjunction with the Gaussian approximation for the multiuser interference were used to obtain a computationally tractable algorithm. Blind multiuser channel estimation has also been addressed in several papers [103,104], with an assumption of coarse synchronization.

Most of the current work, with a few exceptions [105–108], assume square pulse shaping waveforms leading to closed-form optimistic results; see Ref. 107 for a detailed discussion. Furthermore, very little attention has been paid to carrier offset recovery in a multiuser system, except for the results reported in the paper [87]. In this section, we will only discuss channel estimation at the base-station, assuming coarse synchronization and perfect downconversion. For handset channel estimation algorithms, the reader is referred elsewhere [107,109]. Additionally, we restrict our attention to only training-based methods; blind techniques are rarely used in wireless systems.[14] The channel model assuming T training symbols for each user can be written as

$$\mathbf{z} = \mathbf{U}\mathbf{a} + \boldsymbol{\nu} \tag{30}$$

where the size of the vectors $\mathbf{z}$ and $\boldsymbol{\nu}$, and matrix $\mathbf{U}$ is appropriately redefined for an observation length of T symbols, using the definition in (26). The matrix $\mathbf{U}$ depends on the spreading codes, $\phi_i[n]$ and the training symbols $b_i[k]$, all of which are assumed known for all users. Thus, the matrix $\mathbf{U}$ is completely known. The maximum-likelihood estimate of the channel coefficients, $\mathbf{a}$, is given by the pseudoinverse [4,91]:

$$\hat{\mathbf{a}} = (\mathbf{U}^H\mathbf{U})^{-1}\mathbf{U}^H\mathbf{z}. \tag{31}$$

This solution retains several desirable statistical properties of the maximum-likelihood estimates for linear Gaussian problems [110], namely, consistency, unbiasedness and efficiency. Note that there are several leading zeros in $\mathbf{a}$. The variance of the maximum-likelihood estimator $\hat{\mathbf{a}}$ can be reduced by detecting the unknown number of leading (and possibly trailing) zeros in $\mathbf{a}$, which reduces the number of estimated parameters. The above channel estimation procedure can also easily be extended to long-code DS-CDMA systems [111]. In practice the additive noise $\boldsymbol{\nu}$ is better modeled as colored Gaussian noise with unknown covariance due to out-of-cell multiuser interference. The maximum-likelihood estimate of $\mathbf{a}$ requires estimation of the unknown covariance, thereby leading to more accurate results compared to (31) at the expense of increased computation [89].

[14] A notable exception is high-definition television (HDTV) transmission, where no resources are wasted in training symbols, and slow channel time variation permit the use of blind estimation techniques.

Having estimated the channel for all the users, the channel estimates are then used to detect the rest of the information bearing bits in the packet. For bit detection, the received signal representation in (29) is more appropriate, where the matrix $\mathbf{Q}$ is formed using the channel estimates $\hat{\mathbf{a}}$ and the user signature waveforms $\psi_i[n]$.

5.2.3. Multiuser Detection. As a result of channel-induced imperfections and time-varying asynchronism between the users, it is practically impossible to maintain orthogonality between the user signals. *Multiple-access interference* (MAI) is caused by the simultaneous transmission of multiple users, and is the major factor that limits the capacity and performance of DS-CDMA systems. In the second generation CDMA standards, the multiple-access interference is treated as part of the background noise and single-user optimal detection strategy is used. The single-user receiver is prone to the *near–far* problem, where a high-power user can completely drown the signal of a weak user. To avoid the near–far problem, CDMA-based IS-95 standard uses tight power control to ensure that all users have equal received power. Even with the equal received power, the output of the single-user detector is contaminated with MAI and is suppressed by using very strong forward error correcting codes.

The MAI is much more structured than white noise, and this structure was exploited [112] to derive the optimal detector that minimizes the probability of error. The optimal detector alleviates the near–far problem that plagues the single-user receiver. The optimal detector, thus, does not require fast power control to achieve a desired level of performance, thereby reducing the system overhead greatly. Further, as the number of users increases, the optimal receiver achieves significant gains over single-user receivers, even with perfect power control. Unfortunately, the optimal receiver is computationally too complex to be implemented for large systems [113]. The computational intractability of multiuser detection has spurred a rich literature on developing low-complexity suboptimal multiuser detectors.

Most of the proposed suboptimal detectors can be classified in one of two categories: linear multiuser detectors and subtractive interference cancellation detectors. Linear multiuser receivers linearly map the soft outputs of single-user receivers to an alternate set of statistics, which can possibly be used for an improved detection. In subtractive interference cancellation, estimates for different user signals are generated and then removed from the original signal.

To gain insight into different methods for multiuser detection, we will limit the discussion in this section to a simple case of no multipath and no carrier frequency errors. We further assume that the pulse-shaping introduces no ISI and all users are synchronous, thereby leading to simplification of (29) as

$$\mathbf{z}[k] = \mathbf{Q}\mathbf{b}[k] + \boldsymbol{\nu}[k] \tag{32}$$

where $\mathbf{Q} = [\mathbf{q}_1\ \mathbf{q}_2\ \cdots\ \mathbf{q}_K]$, $\mathbf{q}_i = [q_i[1]\ q_i[2]\ \cdots\ q_i[NP]]^T$, and $\mathbf{b}[k] = [b_1[k]\ b_2[k]\ \cdots\ b_K[k]]^T$. Note that this simplification

only eliminates ISI, not the multiple access interference, which is the primary emphasis of the multiuser detection. We quickly note that all the subsequently discussed multiuser detection methods can be extended to the case of asynchronous and ISI channels. The code matched-filter outputs, $\mathbf{y}[k] = \mathbf{Q}^H\mathbf{z}[k]$ can be written as

$$\mathbf{y}[k] = \mathbf{R}\mathbf{b}[k] + \boldsymbol{\nu}[k] \tag{33}$$

The $K \times K$ matrix $\mathbf{R} = \mathbf{Q}^H\mathbf{Q}$ is the correlation matrix, whose entries are the values proportional to the correlations between all pairs of spreading codes. The matrix $\mathbf{R}$ can be split into two parts, $\mathbf{R} = \mathbf{D} + \mathbf{O}$, where $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{ii} = \mathcal{P}_i$. Thus (33) can be written as follows:

$$\mathbf{y}[k] = \mathbf{D}\mathbf{b}[k] + \mathbf{O}\mathbf{b}[k] + \boldsymbol{\nu}[k] \tag{34}$$

The matrix $\mathbf{O}$ contains the off-diagonal elements of $\mathbf{R}$, with entries proportional to the cross-correlations between different user codes. The first term in (34), $\mathbf{b}[k]$, is simply the decoupled data of each user and the second term, $\mathbf{O}\mathbf{b}[k]$, represents the MAI.

5.2.3.1. Matched-Filter Detector. Also known as *single-user optimal receiver*, the matched-filter receiver treats the MAI + $\boldsymbol{\nu}[k]$ as white Gaussian noise, and the bit decisions are made by using the matched-filter outputs, $\mathbf{y}[k]$. The hard bit decisions are made as

$$\hat{\mathbf{b}}_{\mathrm{MF}}[k] = \mathrm{sign}(\mathbf{y}[k]) \tag{35}$$

where sign$(\cdot)$ is a nonlinear decision device and outputs the sign of the input. The matched-filter receiver is extremely simple to implement and requires no knowledge of MAI for its implementation. However, the matched-filter receiver suffers from the near–far problem, where a nonorthogonal strong user can completely overwhelm a weaker user; in fading environments, power disparities are commonly encountered and perfect power control is generally impossible.

5.2.3.2. Maximum A Posteriori Probability (MAP) Detector. As the name suggests, the maximum–likelihood detector chooses the most probable sequence of bits to maximize the joint a posteriori probability, the probability that particular bits were transmitted having received the current signal: Prob($\mathbf{b}[k]|\mathbf{r}(t)$, for all t). The MAP detector minimizes the probability of error [112]. Under the assumption that all bits are equally likely, the MAP detector is equivalent to the maximum–likelihood detector, which finds the bits $\mathbf{b}[k]$ that maximize the probability Prob($\mathbf{r}(t)|\mathbf{b}[k]$).

For the case of K synchronous users in (32), there are 2^K possible transmitted bit combinations in each received symbol duration. Thus, the computation of the maximum–likelihood bit estimates requires number of operations proportional to 2^K. For large number of users, the number of operations to obtain maximum–likelihood estimates become prohibitive for real-time implementation.

In the general case of asynchronous users, if a block of $M \leq G$ bits per user is used to perform the detection, there are 2^{MK} possible bit decisions, $\{\mathbf{b}[k]\}_{k=1}^{M}$. An exhaustive search over all possible bit combinations is clearly impractical, even for moderate values of M and K. However, the maximum-likelihood detector can be implemented using the Viterbi algorithm [114]; the Viterbi implementation (see Section 5.2.4 for more details on Viterbi decoding) is similar to maximum-likelihood sequence detection for ISI channels [4]. The resulting Viterbi algorithm has a complexity that is linear in block length M and exponential in the number of users, of the order of $M2^K$.

The maximum-likelihood detector requires complete knowledge of all user parameters that include not only the spreading signatures of all users but also their channel parameters. The channel parameters are unknown a priori, and have to be estimated. Despite the huge performance and capacity gains of the maximum-likelihood detector, it remains impractical for real-time systems. The computational intractability of the ML detector has led to several detectors which are amenable to real-time implementation.

5.2.3.3. Linear Detectors. Linear detectors map the matched filter outputs, $\mathbf{y}[k]$, in Eqn. (33) into another set of statistics to reduce the MAI experienced by each user. Two of the most popularly studied matched-filter receivers are the decorrelating detector and minimum mean-squared error (MMSE) detector.

The **decorrelating detector** was proposed in 1979 and 1983 [115,116] and later analyzed [117,118]. The decorrelating detector uses the inverse of the correlation matrix, $\mathbf{R}^{-1}$, to decouple the data of different users. The output of the decorrelating detector before hard decision is given by

$$\hat{\mathbf{b}}_{dec}[k] = \mathbf{R}^{-1}\mathbf{y}[k] \tag{36}$$

$$= \mathbf{b}[k] + \mathbf{R}^{-1}\boldsymbol{\nu}[k] \tag{37}$$

$$= \mathbf{b}[k] + \boldsymbol{\nu}_{dec}[k] \tag{38}$$

The decorrelating detector completely suppresses the MAI at the expense of reduced signal power.[15] For nonmultipath channels and unknown user amplitudes, the decorrelating detector yields optimal maximum-likelihood estimates of the bits and the received amplitudes. The decorrelating detector leads to substantial performance improvements over the single-user detector [118] if the background noise is low compared to the MAI. In addition to the noise enhancement problem, the computational complexity of the decorrelating detector can be prohibitive to implement in real-time; however, dedicated application-specific integrated circuits (ASICs) can ameliorate the real-time implementation issues. The computational complexity of the decorrelating detector prohibits its use for long-code CDMA systems, since it requires recomputation of $\mathbf{R}^{-1}$ for every bit.

The **MMSE detector** [119] accounts for the background noise and the differences in user powers to suppress the MAI. The detector is designed to minimize the

[15] The decorrelating detector is very similar to the zero-forcing equalizer [4], which is used to completely suppress ISI.

mean-squared error between the actual data, $\mathbf{b}$ and the soft estimate of data, $\hat{\mathbf{b}}_{\text{mmse}}$. The MMSE detector hard limits the following transform of the received signal:

$$\hat{\mathbf{b}}_{\text{mmse}} = (\mathbf{R} + \sigma^2\mathbf{I})^{-1}\,\mathbf{y}[k]. \tag{39}$$

The MMSE detector[16] balances between the suppression of MAI and suppression of background noise. The higher the background noise level, the lesser is the emphasis on suppressing MAI and vice versa. The MMSE detector has been shown to have a better probability of error than the decorrelating detector [12]. It is clear that as the background noise goes to zero, the MMSE detector converges to the decorrelating detector. On the other hand, as the background noise becomes more dominant compared to MAI, the MMSE detector converges to a single-user detector. Unlike the decorrelator and single-user receiver, the MMSE detector requires an estimate of user amplitudes. Further, the complexity of the MMSE detector is similar to that of the decorrelator.

A blind extension of the MMSE detector, which does not require the knowledge of other user codes and parameters, has been presented [120]. The blind MMSE is similar to the commonly used beamformer in antenna array processing [121]. The probability of error performance of the MMSE detector was studied [122]. The MMSE estimator was extended to multiple data rate systems, like the third-generation standards [123,124].

5.2.3.4. Subtractive Interference Cancellation. The basic idea in subtractive interference cancellation is to separately estimate the MAI contribution of each user and use the estimates to cancel a part or all the MAI seen by each user. Such a detector structure can be implemented in multiple stages, where each additional stage is expected to improve the accuracy of the decisions. The bit decisions used to estimate MAI can be hard (after the sign$(\cdot)$ operation) or soft (before the sign$(\cdot)$ operation). The nonlinear hard-decision approach uses the bit decisions and the amplitude estimates of each user to estimate the MAI. In the absence of reliable estimates, the hard-decision detectors may perform poorly as compared to their soft-decision counterparts [125,126].

The **successive interference cancellation** (SIC) detector cancels interference serially. At each stage of the detector, bit decisions are used to regenerate a user signal and cancel out the signal of one additional user from the received signal. After each cancellation, the rest of the users see a reduced interference. The SIC detector is initialized by ranking all the users by their received power. For the following discussion, assume that the subscripts represent the user rank based on their received powers. The received signal corresponding to user 1 is denoted by $z_1[n]$ [cf. (32)], and its bit estimate is denoted by $b_1[n]$. The SIC detector includes the following steps:

1. Detect the strongest user bit, $b_1[k]$, using the matched-filter receiver.
2. Generate an estimate, $\hat{z}_1[n]$, of the user signal based on the bit estimate, $b_1[k]$, and the channel estimate.
3. Subtract $\hat{z}_1[n]$ from the received signal $z[n]$, yielding a signal with potentially lower MAI.
4. Repeat steps 1–3 for each of the successive users using the "cleaned" version of the signal from the previous stage.

Instead of using the hard bit estimates, $\hat{b}_i[k]$, soft bit estimates (without the sign operator) can also be used in step 3. If reliable channel estimates are available, hard-decision SIC generally outperforms the soft-decision SIC; the situation may reverse if the channel estimates have poor accuracy [125,126]. The reasons for canceling the signals in descending order of received signal strength are as follows: (1) acquisition of the strongest user is the easiest and has the highest probability of correct detection; (2) removal of the strongest user greatly facilitates detection of the weaker users — the strongest user sees little or no interference suppression, but the weakest user can potentially experience a huge reduction in MAI; and (3) SIC is information-theoretically optimal, that is, optimal performance can be achieved using SIC [127].

The SIC detector can improve the performance of the matched-filter receiver with minimal amount of additional hardware, but SIC presents some implementation challenges: (1) each stage introduces an additional bit delay, which implies that there is a tradeoff between the maximum number of users that are canceled and the maximum tolerable delay [128]; and (2) time variation in the received powers caused by time-varying fading requires frequent reordering of the signals [128]. Again, a tradeoff between the precision of the power ordering and the acceptable processing complexity has to be made.

Note that the performance of SIC is dependent on the performance of the single-user matched filter for the strongest users. If the bit estimates of the strongest users are not reliable, then the interference due to the stronger users is quadrupled in power (twice the original amplitude implies 4 times the original power). Thus, the errors in initial estimates can lead to large interference power for the weaker users, thereby amplifying the near-far effect. So, for SIC to yield improvement over the matched filter, a certain minimum performance level of the matched-filter is required.

In contrast to the SIC detector, the **parallel interference cancellation** (PIC) detector [129] estimates and cancels MAI for all the users in parallel. The PIC detector is also implemented in multiple stages:

1. The first stage of the PIC uses a matched-filter receiver to generate bit estimates for all the users, $\hat{\mathbf{b}}_{\text{MF}}[k]$.
2. The signal for the matched filter for user i in the next stage is generated as follows. Using the effective spreading codes and the bit estimates of all except the ith user, the MAI for user i is generated and subtracted from the received signal, $r[n]$.
3. The signal with canceled MAI is then passed to the next stage, which hopefully yields better bit estimates.

[16] The MMSE detector is similar to the MMSE linear equalizer used to suppress ISI [4].

4. Steps 1–3 can be repeated for multiple stages. Each stage uses the data from the previous stage and produces new bit estimates as its output.

The output of $(m+1)$st stage of the PIC detector can be concisely represented as

$$\hat{\mathbf{b}}^{(m+1)}[k] = \text{sign}\,(\mathbf{y}[k] - \mathbf{O}\hat{\mathbf{b}}^{(m)}[k]) = \text{sign}\,(\mathbf{D}\mathbf{b}[k] + \mathbf{O}(\mathbf{b}[k] - \hat{\mathbf{b}}^{(m)}[k]) + \boldsymbol{\nu}[k]) \quad (40)$$

The term $\mathbf{O}\hat{\mathbf{b}}^{(m)}[k]$ is the estimate of MAI after the mth stage. Since soft-decision SIC exploits power variation by canceling in the order of signal strength, it is superior in a non-power-controlled system. On the other hand, soft-decision PIC has a better performance in a power-controlled environment. Performance evaluation of soft-decision PIC can be found elsewhere [130,131], as well as comparison of the soft-decision PIC and SIC detectors [130].

The susceptibility of the PIC to the initial bit estimates has been discussed [129]. An improved PIC scheme, which uses a decorrelator in the first stage, has been proposed [132]. The decorrelator-based PIC detector provides significant performance gains over the original PIC scheme. Further improvements to PIC detector's performance can be obtained by linearly combining the outputs of different stages of the detector [133].

For long-code systems, multistage detection is best suited for its good performance–complexity tradeoff. Multistage detection requires only matrix multiplications in each processing window while other multiuser detectors such as the decorrelator and MMSE detector require matrix inversions during each processing window due to the time-varying nature of the spreading codes.

5.2.4. Channel Decoding. Following the multiuser detection, the detected symbols are decoded using a channel decoder to produce an estimate of the transmitted information bits. In this section we will review decoders for FEC coding when the sender uses either one or more than one transmit antenna. For single-antenna systems, we will consider Viterbi decoding [134] of convolutional codes and review its lower complexity approximations. For multiple antennas, the ML decoder for the Alamouti scheme is presented along with a discussion on complexity of decoding space-time trellis codes.

5.2.4.1. Single-Transmit Antenna. The detected bits after the multiuser detection can be treated to be free of multiple-access interference, and hence a single-user channel decoder can be used. Viterbi decoding for convolutional codes is an application of the dynamic programming principle, and allows efficient hard- or soft-decision decoding of convolutional codes. Furthermore, Viterbi decoding is amenable to VLSI implementation.

To understand the decoding of a convolutional code, an alternate representation for the encoding process, known as a *trellis diagram*, is better suited. A convolutional code is a finite-state machine, whose next state and output are completely determined by its current state and input. The states of a convolutional code can be depicted using a

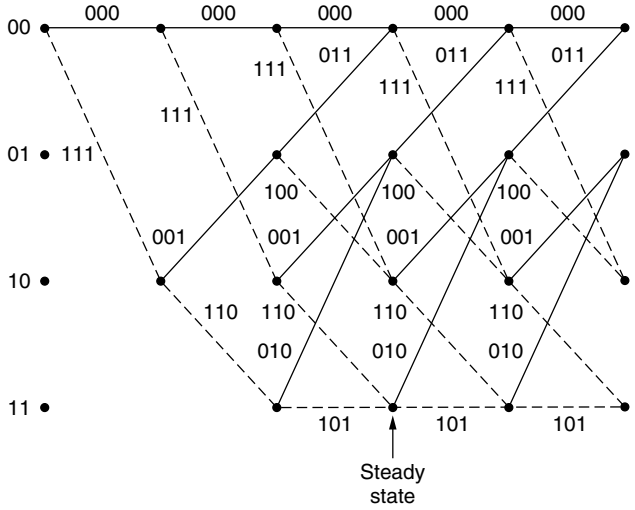


Figure 11. Trellis diagram for the example (3,1) convolutional code.

trellis diagram. The trellis diagram for the example code in Fig. 7 is given in Fig. 11. A close examination of the trellis diagram in Fig. 11 reveals that the diagram repeats itself after three stages, which is equal to the constraint length of the code, $S = 3$. In fact, the three outputs are completely determined by the first two states of the system and the input, which explains the four possible states (00, 01, 10, and 11) in the trellis and the two possible transitions from the current state to the next state based on the input (0 or 1). The solid transitions are due to input 0, and the dashed line shows transition due to input 1. The numbers along the transition describe the output of the decoder due to that transition.

Assume that κ encoded bits were sent using a rate R convolutional code; note that κ can be less than the packet length G if a training sequence is sent in the packet for channel estimation. The maximum a posteriori decoder chooses the information bit sequence that maximizes the posterior probability of the transmitted information symbols given the received noise corrupted signal. To compute the exact estimate of the transmitted information symbols, a total of $2^{\kappa R}$ bits should be considered. It was shown that, as a result of the encoding structure of the convolutional codes, the optimal decoder has a complexity which is linear in the codeword length κ [134]. In the Viterbi algorithm, a metric is associated with each branch of the code trellis. The metric associated with a branch at a particular stage or level i, is the probability of receiving r_i, when the output corresponding to that branch is transmitted. A *path* is defined as a sequence of branches at consecutive levels so that the terminal node of a branch ends in the source node of the next branch. The metric associated with a path is the sum of the metrics associated with the branches in the path. And the metric associated to a node is the minimum metric associated with any path starting from the start node to that node. With these associations, the MAP codeword corresponds to the path that has the lowest metric from the start node to the final node. If the decoder starts and ends in state 0, with start level labeled 0 and end level labeled κ, then for

all $0 < l < \kappa$, the defining equation in the optimization problem is

$$metric(0, \kappa) = \min_{m \in \text{states}} (metric(0, l_m) + metric(l_m, G)) \quad (41)$$

where $metric(i, j)$ is the minimum metric of any path originating from node i and ending in node j and l_m represent the mth node in level l. With additive Gaussian noise, the metric for each branch is the mean-squared error between the symbol estimate and the received data. Once we know the metric associated with all the nodes in level l, the metric associated with the mth node in level $l + 1$ can be calculated by

$$metric(0, (l+1)_m) = \min_{i \in \text{states}} (metric(0, l_i) + metric(l_i, (l+1)_m)) \quad (42)$$

If there is no branch between the node i in state l and node m in state $l + 1$, then the metric associated with that branch is assumed to be infinitely large.

This iterative method of calculating the optimal code reduces the complexity of the decoder to be linear in codeword length κ. However, at every stage of the trellis, the Viterbi algorithm requires computation of the likelihood of each state. The number of states is exponential in the size of the constraint length, S, of the code, thereby making the total complexity of the algorithm of the order of $\kappa 2^{(S+1)}$.

For large constraint lengths, the Viterbi decoding can be impractical for real-time low-power applications. As applications require higher data rates with increasing reliability, higher constraint lengths are desirable. There have been several low-complexity alternatives to Viterbi decoding proposed in the literature: sequential decoding [135], majority logic decoding [136], M algorithm or list decoding [137,138], T algorithm [139], reduced-state sequence detection [140,141], and maximal weight decoding [59].

As noted in the beginning of this section, most of the channel coding and decoding procedures are designed for single-user AWGN channels or fading channels. In the presence of multiaccess interference, joint multiuser detection and decoding [142–146] can lead to lower error performance at the expense of increased receiver complexity.

5.2.4.2. Multiple Transmit Antennas. The information symbols encoded using the Alamouti scheme in Fig. 8 admit a simple maximum-likelihood decoder. With two transmit and single receive antenna, the sampled received signal in two consecutive time symbols is given by

$$z[1] = h_1 s_1 + h_2 s_2 + n_1$$
$$z[2] = -h_1 s_2^* + h_2 s_1^* + n_1$$

where n_1 and n_2 are assumed to be independent instances of circularly symmetric Gaussian noise with zero mean and unit variance. The maximum-likelihood detector builds the following two signals:

$$\hat{s}_1 = h_1^* z[1] + h_2 z^*[2] = (|h_1|^2 + |h_2|^2) s_1 + h_1^* n_1 + h_2 n_2^*$$
$$\hat{s}_2 = h_2^* z[1] - h_1 z^*[2] = (|h_1|^2 + |h_2|^2) s_2 - h_0 n_1^* + h_1^* n_1 \quad (43)$$

followed by the maximum-likelihood detector for each symbol s_i, $i = 1, 2$. The combined signals in (43) are equivalent to that obtained from a two-branch receive diversity using maximal ratio combining (MRC) [6]. Thus, the Alamouti scheme provides an order two transmit diversity much like an order two receive diversity using MRC. Note that both the Alamouti and MRC schemes have the same average transmission rate, one symbol per transmission, but the Alamouti scheme requires at least two transmissions to achieve order two diversity, while MRC achieves order two diversity per transmission.

If a space-time trellis code is used, then the decoder is a simple extension of the decoder for the single-antenna case. As the number of antennas is increased to achieve higher data rates, the decoding complexity increases exponentially in the number of transmit antennas [5], thereby requiring power-hungry processing at the receiver. Though there is no work on reduced complexity decoders for space-time trellis codes, complexity reduction concepts for single-antenna trellis decoding should apply (see text above).

5.3. Power Control

Power control was amply motivated on the capacity grounds in Sections 4.1 and 4.2; in this section, we will only highlight some of the representative research on power control methods and its benefits. Power control is widely used in second- and third-generation cellular systems. For instance, in IS-95, transmit power is controlled not only to counter the near–far effect but also to overcome the time-varying fading. By varying the transmit power based on the channel conditions, a fixed received signal-to-noise ratio (SNR) can be achieved. A SNR guarantee implies a guarantee on the reliability of received information, through the relation between the packet error rate and the received SNR [4].

Information-theoretically optimal power control for a multiuser system was discussed elsewhere [147–150]. While providing a bound on the achievable capacity, the proposed power control algorithms assume perfect knowledge of the time-varying channel at the transmitter. Hence, the power control policies and the resultant system performance is only a loose bound for the achievable performance. Network capacity analysis with power control errors has appeared in Refs. 151,152, and references therein.

Significant research effort has been devoted to power control algorithms for data traffic [e.g., 153–160]. Most of the above work on power control has been for circuit-switched networks, where users are given a certain dedicated channel for their entire session. With the advent of services supporting bursty traffic, such as email and Web browsing, resource allocation for shared channels and packet networks becomes of importance. First steps in these directions can be found in the literature [158,159,161]. Lastly, we note that power control can also lead to gain in packet-switched networks, like IEEE 802.11 or ad hoc networks; preliminary results can be found elsewhere [162,163].

6. CONCLUSIONS

If the relentless advances in wireless communications since 1990 are an indicator of things to come, then it is clear that we will witness not only faster ways to communicate but also newer modes of communication. The fundamental information theoretic bounds hold as long as the assumed communication model holds. The capacity of the channel can be "increased," by introducing new capabilities such as multiple antennas and ad hoc networking. Thus, it will be safe to conclude that the actual physical limits of wireless communication are still unknown and it is for us to exploit that untapped potential with a mix of creativity and serendipity.

BIOGRAPHIES

Ashutosh Sabharwal received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, New Delhi, India, in 1993. He received his M.S. and Ph.D. degrees in electrical engineering in 1995 and 1999, respectively, from the Ohio State University, Columbus, Ohio. Since 1999, he has been a postdoctoral research associate at the Center for Multimedia Communication, Rice University, Houson, Texas, where he currently is a faculty fellow. He was the recipient of the 1999 Presidential Dissertation Fellowship sponsored by Ameritech. His current research interests include wireless communications, network protocols, and information theory.

Behnaam Aazhang received his B.S. (with highest honors), M.S., and Ph.D. degrees in electrical and computer engineering from the University of Illinois at Urbana–Champaign in 1981, 1983, and 1986, respectively. From 1981, to 1985, he was a research assistant in the Coordinated Science Laboratory at the University of Illinois. In August 1985, he joined the faculty of Rice University, Houston, Texas, where he is now the J. S. Abercrombic Professor in the Department of Electrical and Computer Engineering and the Director of Center for Multimedia Communications. He has been a Visiting Professor at IBM Federal Syttems Company, Houston, Texas; and Laboratory for Communication Technology at Swiss Federal Insitute of Technology (ETH), Zurich, Switzerland; the Telexommunications Laboratory at University of Oulu, Oulu, Finland; and the U.S. Air Force Phillips Laboratory, Albuquerque, New Mexico. His research interests are in the arcas of communication theory, information theory, and their applications with emphasis on multiple access communications, cellular mobile radio communications, and optical communication networks. Dr. Aazhang is a Fellow is IEEE, a recipient of the Alcoa Foundation Award 1993, the NSF Engineering Initiation Award 1987–1989, and the IBM Graduate Fellowship 1984–1985, and is a member of Tau Beta Pi and Eta Kappa Nu. He currently is serving on Houston Mayor's Commission on Cellular Towers. He has served as the editor for spread Spectrum Networks of *IEEE Transactions on Communications* 1993–1998; the treasurer of the IEEE Information Theory Society 1995–1998; the technical area chair of the 1997 Asilomar Conference, Monterey, California; the secretary of the Information Theory Society 1990–1993; the publications chairman of the 1993 IEEE International Symposium on Information Theory, San Antonio, Texas; the co-chair of the Technical Program Committee of 2001 Multi-Dimensional and Mobile Communication (MDMC) Conference in Pori, Finland.

BIBLIOGRAPHY

1. R. Steele, J. Whitehead, and W. C. Wong, System aspects of cellular radio, *IEEE Commun. Mag.* **33**: 80–86 (Jan. 1995).
2. U. T. Black, *Mobile and Wireless Networks*, Prentice-Hall, 1996.
3. J. Geier, *Wireless LANs: Implementing Interoperable Networks*, Macmillan Technical Publishing, 1998.
4. J. G. Proakis, *Digital Communications*, McGraw-Hill, 1995.
5. V. Tarokh, N. Seshadri, and A. R. Calderbank, Space-time codes for high data rate wireless communication: Performance criterion and code construction, *IEEE Trans. Inform. Theory* **44**: 744–765 (March 1998).
6. D. G. Brennan, Linear diversity combining techniques, *Proc. IRE*, 1959.
7. T. S. Rappaport, *Wireless Communications: Principles and Practice*, Prentice-Hall, 1996.
8. M. G. Jansen and R. Prasad, Capacity, throughput, and delay analysis of a cellular DS-CDMA system with imperfect power control and imperfect sectorization, *IEEE Trans. Vehic. Technol.* **44**: 67–75 (Feb. 1995).
9. A. Sabharwal, D. Avidor, and L. Potter, Sector beam synthesis for cellular systems using phased antenna arrays, *IEEE Trans. Vehic. Technol.* **49**: 1784–1792 (Sept. 2000).
10. E. S. Sousa, V. M. Jovanović, and C. Daigneault, Delay spread measurements for the digital cellular channel in Toronto, *IEEE Trans. Vehic. Technol.* **43**: 837–847 (Nov. 1994).
11. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, 1991.
12. S. Verdú, *Multiuser Detection*, Cambridge Univ. Press, 1998.
13. S. Verdú and S. Shamai (Shitz), Spectral efficiency of CDMA with random spreading, *IEEE Trans. Inform. Theory* **45**: 622–640 (March 1999).
14. S. Shamai (Shitz) and S. Verdú, The impact of frequency flat fading on the spectral efficiency of CDMA, *IEEE Trans. Inform. Theory* **47**: 1302–1327 (May 2001).
15. I. Katzela and M. Nagshineh, Channel assignment schemes for cellular mobile telecommunication systems: A comprehensive survey, *IEEE Pers. Commun.* 10–31 (June 1996).
16. S. Jordan, Resource allocation in wireless networks, *J. High Speed Networks* **5**(1): 23–24 (1996).
17. C. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error-correcting coding and decoding: Turbo codes, *Proc. 1993 Int. Conf. Communications*, Geneva, Switzerland, May 1993, pp. 1064–1070.
18. P. A. Bello, Characterization of randomly time-variant linear channels, *IEEE Trans. Commun. Syst.* **CS-11**: 360–393 (Dec. 1963).
19. E. Biglieri, J. Proakis, and S. Shamai, Fading channels: Information-theoretic and communication aspects, *IEEE Trans. Inform. Theory* **44**: 2619–2692 (Oct. 1998).

20. E. Malkamäki and H. Leib, Coded diversity on block-fading channels, *IEEE Trans. Inform. Theory* **45**: 771–781 (March 1999).

21. T. L. Marzetta and B. M. Hochwald, Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading, *IEEE Trans. Inform. Theory* **45**(1): 139–157 (1999).

22. R. Knopp and P. A. Humblet, On coding for block fading channels, *IEEE Trans Inform. Theory* **46**: 189–205 (Jan. 2000).

23. C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* **27**: 379–423 (Part I), 623–656 (Part II) (1948).

24. L. H. Ozarow, S. Shamai, and A. D. Wyner, Information theoretic considerations for cellular mobile radio, *IEEE Trans. Inform. Theory* **43**: 359–378 (May 1994).

25. I. E. Telatar, *Capacity of Multi-Antenna Gaussian Channels*, Technical Report, AT&T Bell Labs, 1995; [appeared in *Eur. Trans. Telecommun.* **10**(6): 585–595 (1999)].

26. A. J. Goldsmith and P. P. Varaiya, Capacity of fading channels with channel side information, *IEEE Trans. Inform. Theory* **43**: 1986–1992 (Nov. 1997).

27. G. Caire, G. Taricco, and E. Biglieri, Optimum power control over fading channels, *IEEE Trans. Inform. Theory* **45**: 1468–1489 (July 1999).

28. A. Sabharwal, E. Erkip, and B. Aazhang, On side information in multiple antenna block fading channels, *Proc. ISITA*, Honolulu, Hawaii, Nov. 2000.

29. A. Narula, M. J. Lopez, M. D. Trott, and G. W. Wornell, Efficient use of side information in multiple-antenna data transmission over fading channels, *IEEE-JSAC* **16**: 1423–1436 (Oct. 1998).

30. A. Narula, M. D. Trott, and G. W. Wornell, Performance limits of coded diversity methods for transmitter antenna arrays, *IEEE Trans. Inform. Theory* **45**: 2418–2433 (Nov. 1999).

31. J.-C. Guey, M. Fitz, M. Bell, and W. Y. Kuo, Signal design for transmitter diversity wireless communication systems over raleigh fading channels, *IEEE Trans. Commun.* **46**: 527–537 (April 1999).

32. V. Tarokh, A. Naguib, N. Seshadri, and A. R. Calderbank, Space-time codes for high data rate wireless communication: Performance criteria in the presence of channel estimation errors, mobility, and multiple paths, *IEEE Trans. Commun.* **47**: 199–207 (Feb. 1999).

33. A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill International Editions, 1984.

34. S. Verdú and T. S. Han, A general formula for channel capacity, *IEEE Trans. Inform. Theory* **40**: 1147–1157 (July 1994).

35. H. Viswanathan, *Capacity of Fading Channels with Feedback and Sequential Coding of Correlated Sources*, PhD thesis, Cornell Univ., Aug. 1997.

36. E. Biglieri, G. Caire, and G. Taricco, Limiting performance of block-fading channels with multiple antennas, *IEEE Trans. Inform. Theory* (Aug. 1999).

37. G. J. Foschini, Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas, *Bell Labs Tech. J.* 41–59 (1996).

38. G. Caire and S. Shamai (Shitz), On the capacity of some channels with channel state information, *IEEE Trans. Inform. Theory* **45**: 2007–2019 (Sept. 1999).

39. S. Shamai and A. D. Wyner, Information theoretic considerations for symmetric, cellular, multiple-access fading channels — Part I, *IEEE Trans. Inform. Theory* **43**: 1877–1894 (Nov. 1997).

40. G. W. Wornell, Spread-signature CDMA: Efficient multiuser communication in presence of fading, *IEEE Trans. Inform. Theory* **41**: 1418–1438 (Sept. 1995).

41. R. L. Pickholtz, D. L. Schilling, and L. B. Milstein, Theory of spread-spectrum communications — a tutorial, *IEEE Trans. Commun.* **COM-30**: 855–884 (May 1982).

42. E. Erkip and B. Aazhang, Multiple access schemes over multipath fading channels, *Proc. ISIT*, Cambridge, MA, Aug. 1998.

43. A. D. Wyner, Shannon-theoretic approach to Gaussian cellular multiple access channel, *IEEE Trans. Inform. Theory* **40**: 1713–1727 (Nov. 1994).

44. O. Somekh and S. Shamai (Shitz), Shannon-theoretic approach to a Gaussian cellular multiple-access channel with fading, *Proc. 1998 IEEE Int. Symp. Information Theory*, Cambridge, MA, Aug. 1998, p. 393.

45. A. J. Viterbi, *CDMA: Principles of Spread Spectrum Communication*, Addison-Wesley, 1995.

46. D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, 1992.

47. R. G. Gallager, A perspective on multiaccess channels, *IEEE Trans. Inform. Theory* **IT-31**: 124–142 (March 1985).

48. A. Ephremides and B. Hajek, Information theory and communication networks: An unconsummated union, *IEEE Inform. Theory* **44**: 2416–2434 (Oct. 1998).

49. I. E. Telatar and R. G. Gallager, Combining queuing theory with information theory for multiaccess, *IEEE J. Select. Areas Commun.* **13**: 963–969 (Aug. 1995).

50. M. Win and R. A. Scholz, Impulse radio: How it works, *IEEE Commun. Lett.* **2**: 36–38 (Feb. 1998).

51. K. A. S. Immink, P. H. Siegel, and J. K. Wolf, Codes for digital recorders, *IEEE Trans. Inform. Theory* **44**: 2260–2299 (Oct. 1998).

52. J. D. J. Costello, J. Hagenauer, H. Imai, and S. B. Wicker, Applications of error-control coding, *IEEE Trans. Inform. Theory* **44**: 2531–2560 (Oct. 1998).

53. A. R. Calderbank, The art of signalling, *IEEE Trans. Inform. Theory* **44**: 2561–2595 (Oct. 1998).

54. I. Blake, C. Heegard, T. Høholdt, and V. Wei, Algebraic-geometry codes, *IEEE Trans. Inform. Theory* **44**: 2596–2618 (Oct. 1998).

55. E. R. Berlekamp, *Algebraic Coding Theory*, McGraw-Hill, New York, 1968.

56. G. D. Forney, *Concatenated Codes*, MIT Press, Cambridge, MA, 1966.

57. J. H. van Lint, *Introduction to Coding Theory*, Springer-Verlag, New York, 1992.

58. V. S. Pless, W. C. Huffman, and R. A. Brualdi, eds., *Handbook of Coding Theory*, Elsevier, New York, 1998.

59. S. Das, *Multiuser Information Processing in Wireless Communication*, PhD thesis, Rice Univ., Houston, TX, Sept. 2000.

60. A. Dholakia, *Introduction to Convolutional Codes with Applications*, Kluwer Academic Publishers, 1994.
61. D. Divsalar and M. K. Simon, The design of trellis coded MPSK for fading channels: Performance criteria, *IEEE Trans. Commun.* **36**: 1004–1012 (Sept. 1988).
62. G. Ungerboeck, Channel coding with multilevel/phase signals, *IEEE Trans. Inform. Theory* **28**: 55–67 (Jan. 1982).
63. G. D. Forney and G. Ungerboeck, Modulation and coding for linear Gaussian channels, *IEEE Trans. Inform. Theory* **44**: 2384–2415 (Oct. 1998).
64. J. L. Massey, Coding and modulation in digital communications, *Proc. 1974 Int. Zurich Seminar on Digital Communications*, Zurich, Switzerland, March 1974, pp. E2(1)–E2(4).
65. J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*, Wiley, New York, 1965.
66. R. G. Gallager, *Low Density Parity-Check Codes*, MIT Press, Cambridge, MA, 1962.
67. S. L. Goff, A. Glavieux, and C. Berrou, Turbo-codes and high spectral efficiency modulation, *Proc. 1994 Int. Conf. Communications*, May 1994, Vol. 2, pp. 645–649.
68. W. Liu and S. G. Wilson, Rotationally-invariant concatenated (turbo) TCM codes, *Proc. 1999 Asilomar Conf. Signal System Comp.*, Oct. 1999, Vol. 1, pp. 32–36.
69. O. Y. Takeshita and J. D. J. Costello, New deterministic interleaver designs for turbo codes, *IEEE Trans. Inform. Theory* **46**: 1988–2006 (Sept. 2000).
70. Y. Liu, M. Fitz, and O. Y. Takeshita, QPSK space-time turbo codes, *Proc. 2000 Int. Conf. Communications*, June 2000, Vol. 1, pp. 292–296.
71. V. Tarokh, A. Vardy, and K. Zeger, Universal bound on the performance of lattice codes, *IEEE Trans. Inform. Theory* **45**: 670–681 (March 1999).
72. J. H. Winters, Smart antennas for wireless systems, *IEEE Pers. Commun.* **5**: 23–27 (Feb. 1998).
73. G. J. Foschini and M. J. Gans, On limits of wireless communication in a fading environment when using multiple antennas, in *Wireless Personal Communications*, Kluwer Academic Publishers, 1998.
74. V. Tarokh, H. Jafarkhani, and A. R. Calderbank, Space-time block codes from orthogonal designs, *IEEE Trans. Inform. Theory* **45**: 1456–1467 (July 1999).
75. D. M. Ionescu, New results on space time code design criteria, *Proc. IEEE Wireless Communications and Networking Conf.*, New Orleans, Oct. 1999.
76. O. Tirkkonnen and A. Hottinen, Complex space-time block codes for four Tx antennas, *Proc. GLOBECOM*, 2000 pp. 1005–1009.
77. S. Baro, G. Bauch, and A. Hansmann, Improved codes for space-time trellis coded modulation, *IEEE Commun. Lett.* **4**: 20–22 (Jan. 2000).
78. A. R. Hammons and H. E. Gamal, On the theory of space-time codes for PSK modulation, *IEEE Trans. Inform. Theory* **46**: 524–542 (March 2000).
79. T. Moharemovic and B. Aazhang, Information theoretic optimality of orthogonal space-time transmission schemes and concatenated code construction, *Proc. Int. Conf. Communications* (*ICC*), Acapulco, Mexico, May 2000.
80. M. J. Borran, M. Memarzadeh, and B. Aazhang, Design of coded modulation schemes for orthogonal transmit diversity, *IEEE Trans. Commun.* (in press).
81. S. M. Alamouti, A simple transmit diversity technique for wireless communications, *IEEE J. Select. Areas Commun.* **16**: 1451–1458 (Oct. 1998).
82. R. Prasad, Overview of wireless personal communications: Microwave perspectives, *IEEE Commun. Mag.* 104–108 (April 1997).
83. G. Taricco, E. Biglieri, and G. Caire, Limiting performance of block-fading channels with multiple antennas, *Proc. Inform. Theory Communication Workshop*, pp. 27–29, 1999.
84. *http://www.3gpp.org/*.
85. T. Ojanpera and R. Prasad, eds., *Wideband CDMA for Third Generation Mobile Communications*, Artech House Universal Personal Communications Series, 1998.
86. A. Lapidoth and P. Narayan, Reliable communication under channel uncertainty, *IEEE Trans. Inform. Theory* **44**: 2148–2177 (Oct. 1998).
87. K. Li and H. Liu, Joint channel and carrier offset estimation in CDMA communications, *IEEE Trans. Signal Process.* **47**: 1811–1822 (July 1999).
88. Z. Pi and U. Mitra, Blind delay estimation in multi-rate asynchronous DS-CDMA systems, *IEEE Trans. Commun.* (2000).
89. C. Sengupta, *Algorithms and Architectures for Channel Estimation in Wireless CDMA Communication Systems*, PhD thesis, Rice Univ., Dec. 1998.
90. Z. Tian, K. L. Bell, and H. L. V. Trees, A quadratically constrained decision feedback equalizer for DS-CDMA communication systems, in *IEEE Workshop on Signal Processing Advances in Wireless Communication*, May 1999, pp. 190–193.
91. C. Sengupta, J. Cavallaro, and B. Aazhang, On multipath channel estimation for DS-CDMA systems with multiple sensors, *IEEE Trans. Commun.* **49**: 543–553 (March 2001).
92. D. N. Godard, Self-recovering equalization and carrier tracking of two-dimensional data communication systems, *IEEE Trans. Commun.* **28**: 1867–1875 (Nov. 1980).
93. J. R. Treichler and B. G. Agree, A new approach to multipath correction of constant modulus signals, *IEEE Trans. Acoust. Speech Signal Process.* **31**: 459–472 (April 1983).
94. W. A. Gardner, A new method for channel identification, *IEEE Trans. Commun.* **39**: 813–817 (June 1991).
95. W. A. Gardner, Exploitation of spectral redundancy in cyclostationary signals, *IEEE Signal Process. Mag.* 14–36 (April 1991).
96. L. Tong, G. Xu, and T. Kailath, Blind identification and equalization based on second-order statistics: A time domain approach, *IEEE Trans. Inform. Theory* **40**: 340–349 (March 1994).
97. E. Moulines, P. Duhamel, J.-F. Cardoso, and S. Mayrargue, Subspace methods for the blind identification of multichannel FIR filters, *IEEE Trans. Signal Process.* **43**: 516–525 (Feb. 1995).
98. H. Liu, G. Xu, L. Tong, and T. Kailath, Recent developments in blind channel equalization: From cyclostationarity to subspaces, *Signal Process.* **50**(1–2): 83–99 (1996).
99. S. E. Bensley and B. Aazhang, Subspace-based channel estimation for code division multiple access communications, *IEEE Trans. Commun.* **44**: 1009–1020 (Aug. 1996).

100. E. Ertin, U. Mitra, and S. Siwamogsatham, Maximum-likelihood based multipath channel estimation for code-division multiple-access systems, *IEEE Trans. Commun.* **49**: 290–302 (Feb. 2001).

101. A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Stat. Soc. Ser. B* 1–38 (1977).

102. I. Ziskind and M. Wax, Maximum likelihood localization of multiple sources by alternating projection, *IEEE Trans. Signal Process.* **36**: 1553–1560 (Oct. 1988).

103. M. Torlak and G. Xu, Blind multiuser channel estimation in asynchronous CDMA systems, *IEEE Trans. Signal Process.* **45**: 137–147 (Jan. 1997).

104. M. K. Tsatsanis and G. B. Giannakis, Blind estimation of direct sequence spread spectrum signals in multipath, *IEEE Signal Process.* **45**: 1241–1252 (1997).

105. T. Östman and B. Ottersten, Near far robust time delay estimation for asynchronous DS-CDMA systems with bandlimited pulse shapes, *Proc. IEEE Vehicular Technology Conf.*, May 1998, pp. 1651–1654.

106. V. Tripathi, A. Mantravadi, and V. V. Veeravalli, Channel acquisition for wideband CDMA, *IEEE J. Select. Areas Commun.* **18**: 1483–1494 (Aug. 2000).

107. E. Aktas and U. Mitra, Single user sparse channel acquisition for ds/cdma, *Proc. CISS*, Princeton, NJ, 2000.

108. S. Bhashyam, A. Sabharwal, and U. Mitra, Channel estimation multirate DS-CDMA systems, *Proc. Asilomar Conf. Signal System Comp.*, Pacific Grove, CA, Oct. Nov. 2000.

109. T. P. Krauss and M. D. Zoltowski, Blind channel identification on CDMA forward link based on dual antenna receiver at handset and cross-relation, *Proc. 1999 Asilomar Conf. Signal System Communication*, Oct. 1999, Vol. 1, pp. 75–79.

110. C. R. Rao, *Linear Statistical Inference and Its Applications*, Wiley, New York, 1973.

111. S. Bhashyam and B. Aazhang, Multiuser channel estimation for long code CDMA systems, *Proc. 2000 Wireless Communication Networking Conf.*, 2000.

112. S. Verdú, *Optimum Multiuser Signal Detection*, PhD thesis, Univ. Illinois at Urbana — Champaign, Aug. 1984.

113. S. Verdú, Computational complexity of optimum multiuser detection, *Algorithmica* **4**: 303–312 (1989).

114. S. Verdú, Optimum multiuser asymptotic efficiency, *IEEE Trans. Commun.* **34**: 890–897 (Sept. 1986).

115. K. S. Schneider, Optimum detection of code division multiplexed signals, *IEEE Trans. Aerospace Electron. Syst.* **AES-15**: 181–185 (Jan. 1979).

116. R. Kohno, M. Hatori, and H. Imai, Cancellation techniques of co-channel interference in asynchronous spread spectrum multiple access systems, *Electron. Commun. Japan* **66-A**(5): 20–29 (1983).

117. R. Lupas and S. Verdú, Linear multiuser detectors for synchronous code-division multiple-access channels, *IEEE Trans. Inform. Theory* **35**(1): 123–136 (1989).

118. R. Lupas and S. Verdú, Near-far resistance of multi-user detectors in asynchronous channels, *IEEE Trans. Commun.* **38**: 496–508 (April 1990).

119. Z. Xie, R. T. Short, and C. K. Rushforth, A family of suboptimum detectors for coherent multiuser communications, *IEEE JSAC* **8**: 683–690 (May 1990).

120. M. Honig, U. Madhow, and S. Verdú, Blind adaptive multiuser detection, *IEEE Trans. Inform. Theory* **41**: 944–960 (July 1995).

121. R. Mailloux, *Phased Array Antenna Handbook*, Artech House, 1994.

122. H. V. Poor and S. Verdú, Probability of error in MMSE multiuser detection, *IEEE Inform. Theory* **43**: 858–871 (May 1997).

123. A. Sabharwal, U. Mitra, and R. Moses, Cyclic Wiener filtering based multirate DS-CDMA receivers, *Proc. IEEE WCNC*, New Orleans, Sept. 1999.

124. A. Sabharwal, U. Mitra, and R. Moses, Low complexity MMSE receivers for multirate DS-CDMA systems, *Proc. CISS*. Princeton, NJ, 2000.

125. H. Y. Wu and A. Duel-Hallen, Performance comparison of multi-user detectors with channel estimation for flat Rayleigh fading CDMA channel, *Wireless Pers. Commun.* (July/Aug. 1996).

126. S. D. Gray, M. Kocic, and D. Brady, Multiuser detection in mismatched multiple-access channels, *IEEE Trans. Commun.* **43**: 3080–3089 (Dec. 1995).

127. B. Rimoldi and R. Urbanke, A rate splitting approach to the Gaussian multiple-access channel, *IEEE Trans. Inform. Theory* **42**: 364–375 (March 1996).

128. K. I. Pederson, T. E. Kolding, I. Seskar, and J. M. Holzman, Practical implementation of successive interference cancellation in DS/CDMA systems, *Proc. IEEE Conf. Universal Personal Communication*, 1996, Vol. 1, pp. 321–325.

129. M. K. Varanasi and B. Aazhang, Multistage detection in asynchronous code-division multiple-access communications, *IEEE Trans. Commun.* **38**: 509–519 (April 1990).

130. P. Patel and J. Holzman, Performance comparison of a DS/CDMA system using a successive interference cancellation (IC) scheme and a parallel IC scheme under fading, *Proc. ICC*, New Orleans, May 1994, pp. 510–514.

131. R. M. Buehrer and B. D. Woerner, Analysis of adaptive multistage interference cancellation for CDMA using an improved Gaussian approximation, *Proc. IEEE MILCOM*, San Diego, CA, Nov. 1995, pp. 1195–1199.

132. M. K. Varanasi and B. Aazhang, Near-optimum detection in synchronous code-division multiple-access schemes, *IEEE Trans. Commun.* **39**: 725–736 (May 1991).

133. S. Moshavi, *Multistage Linear Detectors for DS-CDMA Communications*, PhD thesis, City Univ. New York, Jan. 1996.

134. A. J. Viterbi, Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, *IEEE Trans. Inform. Theory* **IT-13**: 260–269 (April 1967).

135. J. M. Wozencraft and B. Reiffen, *Sequential Decoding*, MIT Press, Cambridge, MA, 1961.

136. J. L. Massey, *Threshold Decoding*, MIT Press, Cambridge, MA, 1998.

137. J. B. Anderson and S. Mohan, Sequential coding algorithms: A survey and cost analysis, *IEEE Trans. Commun.* **COM-32**: 169–176 (Feb. 1984).

138. G. J. Pottie and D. P. Taylor, A comparison of reduced complexity decoding algorithms for trellis codes, *IEEE J. Select. Areas Commun.* **7**: 1369–1380 (Dec. 1989).

139. S. T. Simmons, Breadth-first trellis decoding with adaptive effort, *IEEE Trans. Commun.* **38**: 3–12 (Jan. 1990).

140. M. V. Eyuboglu and S. Qureshi, Reduced-state sequence estimation for coded modulation on interference channels, *IEEE J. Select. Areas Commun.* **35**: 944–955 (Sept. 1989).

141. P. R. Chevillat and E. Elephtheriou, Decoding of trellis-encoded signals in the presence of inter-symbol interference and noise, *IEEE Trans. Commun.* **37**: 669–676 (July 1989).

142. C. Schlegel, P. Alexander, and S. Roy, Coded asynchronous CDMA and its efficient detection, *IEEE Trans. Inform. Theory* **44**: 2837–2847 (Nov. 1998).

143. X. Wang and H. V. Poor, Iterative (turbo) soft interference cancellation and decoding for coded CDMA, *IEEE Trans. Commun.* **47**: 1046–1061 (July 1999).

144. H. E. Gamal and E. Geraniotis, Iterative multiuser detection for coded CDMA signals in AWGN and fading channels, *IEEE J. Select. Areas Commun.* **18**: 30–41 (Jan. 2000).

145. R. Chen, X. Wang, and J. S. Liu, Adaptive joint detection and decoding flat-fading channels via mixture Kalman filtering, *IEEE Trans. Inform. Theory* **46**: 2079–2094 (Sept. 2000).

146. L. Wei and H. Qi, Near-optimal limited-search detection on ISI-CDMA channels and decoding of long-convolutional codes, *IEEE Trans. Inform. Theory* **46**: 1459–1482 (July 2000).

147. D. N. C. Tse and S. V. Hanly, Multiaccess fading channels-Part I: polymatroid structure, optimal resource allocation and throughput capacities, *IEEE Trans. Inform. Theory* **44**: 2796–2815 (Nov. 1998).

148. S. V. Hanly and D. N. C. Tse, Multiaccess fading channels — Part II: Delay-limited capacities, *IEEE Trans. Inform. Theory* **44**: 2816–2831 (Nov. 1998).

149. P. Viswanath, V. Anantharam, and D. N. C. Tse, Optimal sequences, power control, and user capacity of synchronous CDMA systems with linear MMSE multiuser receivers, *IEEE Trans. Inform. Theory* **45**: 1968–1983 (Sept. 1999).

150. S. Hanly and D. Tse, Power control and capacity of spread-spectrum wireless networks, *Automatica* **35**: 1987–2012 (Dec. 1999).

151. N. Bambos, Toward power-sensitive network architectures in wireless communications: concepts, issues, and design aspects, *IEEE Pers. Commun.* **5**: 50–59 (June 1998).

152. J. Zhang and E. K. P. Chong, CDMA systems in fading channels: admissibility, network capacity and power control, *IEEE Trans. Inform. Theory* **46**: 962–981 (May 2000).

153. J. Wu and R. Kohno, A wireless multimedia CDMA system based on transmission power control, *IEEE J. Select. Areas Commun.* **14**: 683–691 (May 1996).

154. J. Jacobsmeyer, Congestion relief on power-controlled CDMA networks, *IEEE J. Select. Areas Commun.* **14**: 1758–1761 (Dec. 1996).

155. A. Sampath and J. M. Holtzman, Access control of data in integrated voice/data CDMA systems: benefits and tradeoffs, *IEEE J. Select. Areas Commun.* **15**: 1511–1526 (Oct. 1997).

156. D. Ayyagari and A. Ephremides, Cellular multicode CDMA capacity for integrated (voice and data) services, *IEEE J. Select. Areas Commun.* **17**: 928–938 (May 1999).

157. Y. Lu and R. W. Broderson, Integrating power control, error correction coding, and scheduling for a CDMA downlink system, *IEEE J. Select. Areas Commun.* **17**: 978–989 (May 1999).

158. D. Kim, Rate-regulated power control for supporting flexible transmission in future cdma mobile networks, *IEEE J. Select. Areas Commun.* **17**: 968–977 (May 1999).

159. S. Manji and W. Zhuang, Power control and capacity analysis for a packetized indoor multimedia DS-CDMA network, *IEEE Trans. Vehic. Technol.* **49**: 911–935 (May 2000).

160. D. Goodman and N. Mandayam, Power control for wireless data, *IEEE Pers. Commun.* **7**: 48–54 (April 2000).

161. N. Bambos and S. Kandukuri, Power controlled multiple access (PCMA) in wireless communication networks, *Proc. INFOCOM 2000*, March 2000, Vol. 2, pp. 386–395.

162. P. Gupta and P. R. Kumar, The capacity of wireless networks, *IEEE Trans. Inform. Theory* **46**: 388–404 (March 2000).

163. J. Monks, Power controlled multiple access in *ad hoc* networks, *Proc. Multiaccess, Mobility and Teletraffic for Wireless Communications* (*MMT*), Duck Key, FL, Dec. 2000.

N

NETWORK FLOW CONTROL

STEVEN H. LOW
California Institute of Technology
Pasadena, California

1. INTRODUCTION

Flow and congestion control is a distributed algorithm to share network resources among competing users. Like a transportation network, congestion can build up in a telecommunications network when traffic load exceeds network capacity. When more and more automobiles enter a highway, their speed decreases until everyone is driving at, say, less than 10 km/h instead of 100 km/h, and the network throughput plummets. As load increases in a packet network, queues build up until packets are delayed by an excessive amount or even lost and need to be retransmitted. Packets that are transmitted multiple times or at upstream nodes only to be discarded at downstream nodes, waste network resources and intensify congestion. This has led to congestion collapse where throughput dropped to a small fraction of network capacity [1]. Flow control prevents congestion collapse by adapting user traffic to available capacity.[1]

We distinguish between two types of networks, circuit-switched networks, which we will abbreviate as circuit networks, and packet-switched networks, which we will abbreviate as packet networks. The most important difference between them is that, in a circuit network, when a connection is established between a source and a destination, network resources (e.g., time slots in time-division multiplexed systems, frequency slots in frequency-division multiplexed systems) are reserved along the path for its exclusive use for the duration of the connection. The traditional telephone network is an example of circuit network. The fixed rate allocation simplifies the control of the system and the provisioning of quality of service (QoS). Since network resources (called a "circuit") are dedicated to a connection, they are wasted when the information source of the connection occasionally has no information to send, even if other traffic can make use of these resources. Hence, circuit network is suitable to support applications that generate traffic at a fixed rate, such as uncoded voice. Flow control, that adapts source transmission rate to changes in the availability of network resources along its path, is unnecessary. Traffic is regulated at the connection level through *connection admission control*, which decides whether or not a new connection request is granted, depending on, for example, the availability of resources.

[1] Some authors use *flow* control to refer to mechanisms to avoid a source from overwhelming a receiver and *congestion* control to refer to mechanisms to avoid overloading the network. We make no such distinction and will refer to both as flow control in this article.

In a packet network, in contrast, a path may be established between a source and its destination during the connection setup phase, but no bandwidth or buffer resources are reserved. Rather, these resources are shared by all connections on demand. This is suitable for applications that generate bursty traffic, where periods of activity are interspersed with random idle periods. Sharing of resources dynamically by multiple traffic streams is referred to as *statistical multiplexing*. Statistical multiplexing improves efficiency, since a resource is never idle if there is traffic to be carried, unlike the situation in a circuit network. It, however, makes the control and provisioning of QoS harder. In a circuit network, each connection requires a fixed rate and therefore connection admission control can be easily implemented by checking whether this rate can be supported along the intended path between source and destination. It is difficult to characterize the resource requirements of a bursty source, and hence connection admission control is rarely implemented in packet networks. Since the number of connections in the network is not controlled, the source rates of these connections must be regulated to avoid overwhelming the network or the receiver. This is the purpose of flow control.

In this article, we describe the design objectives (Section 2) and implementation constraints (Section 3) of flow control mechanisms. There is no automatic way to synthesize a flow control scheme that satisfies these objectives, but we can analyze existing or proposed mechanisms through mathematical modeling and computer simulations and apply the understanding to enhance current schemes or design new ones. The goal of this article is to provide an introduction to recent mathematical models for understanding the equilibrium and stability properties of flow control mechanisms (Section 5). Our focus is on general properties that underlie a large class of flow control schemes and therefore many implementation details are abstracted out of the mathematical model. An explanation of network protocols in general is provided in the article [2] in this encyclopedia, and a detailed description of TCP, as well as further references, are provided in the article [3] also in this encyclopedia. To make our discussion concrete, we will use TCP Vegas with DropTail routers, explained in Section 4, for illustration throughout the article.

Our discussion centers around TCP both because it is pervasive — it is estimated that it carries 90% of traffic on the current Internet — and because its distributed, decentralized and asynchronous character allows a scalable implementation. For flow control schemes in asynchronous transfer mode (ATM) networks, see, Ref. 4. A breakthrough in TCP flow control is the algorithm proposed in Ref. 1, which was implemented in the Tahoe version of TCP, and later enhanced into TCP Reno and other variants. These protocols are widely deployed in the current Internet (see [3]). TCP Vegas is proposed in Ref. 5 as an alternative to TCP Reno. Even though it is not widely deployed, it possesses interesting fairness and scalability properties that make it potentially more suitable for

future high speed networks. It also has a simpler analytical structure than Reno that makes it more convenient for us to use as an illustration of the general principles.

In this article, we will use "sources" and "connections" interchangeably. With quantities z_i defined, $z = (z_i) = (z_1, z_2, \ldots)$ denotes the vector whose elements are z_i.

2. DESIGN OBJECTIVES

The objectives of flow control schemes are to share network resources fairly and efficiently, to provide good QoS, and to be responsive to changes in network (or receiver) congestion. What makes the implementation of these objectives challenging is the constraints imposed by decentralization. In this section we elaborate on the design objectives; in the next section we discuss decentralization constraints.

2.1. Fairness

There are many definitions of fairness. Consider a linear network consisting of links $1, 2, \ldots, N$ in tandem, each with a bandwidth capacity of 1 unit. The network is shared by sources $0, 1, 2, \ldots, N$. Source 0 traverses all the N links, while sources $i, i \geq 1$, traverses only link i. If we aim to equally share the bandwidth among the $N + 1$ sources, then each source should get 1/2; this is called *maxmin* fairness [6]. If we aim to maximize the sum of source rates, then each source $i, i \geq 1$, should receive a rate of 1 while source 0 gets 0, to achieve a total source rate of N, almost double that under maxmin fairness if N is large. A compromise, called *proportional* fairness [7], allocates $N/(N + 1)$ to each source $i, i \geq 1$, and $1/(N + 1)$ to source 0.

In general, we can associate a utility function $U_i(x_i)$ with each source i, as a function of the source rate x_i, say, in packets per second. The utility function measures how happy source i is when it transmits at rate x_i. It is usually a concave increasing function, with the interpretation that sources are happier the higher the rate allocation they receive but there is a diminishing return as rate allocation increases. We can then define a rate allocation vector $x = (x_i)$ as *fair*, with respect to utility functions $U = (U_i)$, if it maximizes the aggregate utility $\sum_i U_i$ subject to capacity constraints. For instance, $U_i(x_i) = x_i$ corresponds to maximizing aggregate source rate, and $U_i(x_i) = \log x_i$ corresponds to proportional fairness. We will come back to utility maximization in Section 6.1.

In summary, one of the objectives of flow control is to share network resources fairly, and this can be interpreted as maximizing aggregate utility with different fairness criteria corresponding to different source utility functions.

2.2. Utilization and Quality of Service

Different applications have different quality requirements. For our purposes, we will use QoS to mean packet loss or queueing delay. Packet loss and queueing delay will both be low if the queue length can be kept small.

Recall that packet networks typically do not restrict the number of concurrent sources. If these sources are not flow-controlled, then their aggregate load may exceed the available capacity. Packets will arrive at routers or switches faster than they can be processed and forwarded. Queues will build up, increasing queueing delay, and eventually overflow, leading to packet loss.

Of course, one way to maintain small queues is to underutilize the network, by restricting source rates to (much) less than network capacity. Indeed, if we model the network as an M/M/1 queue, then the model dictates that the input rate must be significantly smaller than the capacity if average queue length is not to be excessive. This suggests an inevitable tradeoff between utilization and QoS: we can achieve either high utilization or high QoS, but *not* both. This view, however, is flawed for it ignores the *feedback* regulation inherent in flow control. It implicitly assumes that the input process remains statistically unchanged as queue builds up indefinitely. With feedback, input rate will be reduced in response to queue buildup, and hence it is possible, with proper control strategy, to stabilize input rate close to the capacity without incurring a large queue.

Ideally, flow control should adapt external traffic load to available capacity, and in equilibrium match the arrival rate to capacity at every bottleneck link. Moreover, queues should then stabilize around a small value, achieving small loss and delay.

As we will explain later, however, it may be difficult to maintain a small queue when congestion information is fed back to traffic sources only implicitly. High utilization and high QoS can both be achieved in a decentralized manner if explicit feedback is available.

2.3. Dynamic Properties

Fairness, utilization, and QoS, such as packet loss and delay, typically are considered as "equilibrium" properties in that usually we only require flow control schemes to achieve these objectives in equilibrium (or stationary regime). Another important criterion by which a flow control scheme is evaluated is its dynamic properties, such as whether the equilibrium point is stable and whether the transition to a new equilibrium is fast.

An ideal flow control scheme should be stable, in the sense that after a disturbance (e.g., arrival or departure of connections), it always converges to a possibly new equilibrium. Moreover, it should converge rapidly, in the presence of network delays, and in an asynchronous environment.

Convergence to equilibrium is desirable because, under a properly designed scheme, fairness, utilization, and QoS objectives are achieved in equilibrium.

2.4. Scalability

Scalability refers to the property that a flow control scheme has a small implementation complexity (implementation scalability) and that it maintains its performance, with respect to the objectives previously discussed (performance scalability), as the network scales up in capacity, propagation delay (geographically reach), and the number of sources.

Flow control schemes that are practical and that are discussed in this article, must be distributed, decentralized, and easy to implement for it not to be a bottleneck itself. The implementation complexity of these schemes typically is low and scalable. Hence, in the rest of the article, we will focus only on performance scalability.

3. INFORMATION CONSTRAINTS

It is important to realize that flow control consists of two algorithms, one carried out by traffic sources to adapt their rates to congestion information on their paths and the other carried out by network resources, often implicitly, to update a measure of congestion whose value is fed back, often implicitly as well, to the sources. On the current Internet, the source algorithm is carried out by TCP and the link algorithm is carried out by a queueing discipline, such as DropTail or RED. Even though the link algorithms are often implicit and overlooked, they are critical in determining the equilibrium and dynamic properties of a network under flow control.

For example, the current TCP uses packet loss as a measure of congestion [3]. A link is considered congested if the loss probability at that link is high. As loss probability increases, a TCP source reduces its rate in response, which in turn causes the link to reduce its loss probability, and so on. The behavior of this feedback loop is determined by how TCP adjusts its rate and how a link implicitly adjusts its congestion measure.

Decentralization requires that the source and the link algorithms use only local information. In this section, we explain the local information available at sources and links for the class of flow control schemes we discuss.

TCP uses "window" flow control, where a destination sends acknowledgments for packets that are correctly received. The time from sending a packet to receiving its acknowledgment is called *round-trip time*. It can be measured at the source and thus does not need clock synchronization between source and destination. A source keeps a variable called window size that limits the number of outstanding packets that have been transmitted but not yet acknowledged. When the window size is exhausted, the source must wait for an acknowledgment before sending a new packet. By numbering the packets and the acknowledgments, the source can estimate the transfer delay of each packet and detect if a packet is lost. Two features are important. The first is the "self-clocking" feature that automatically slows down the source when a network becomes congested and acknowledgments are delayed. The second is that the window size controls the source rate: roughly one window of packets is sent every round-trip time. The first feature was the only congestion control mechanism in the Internet before Jacobson's proposal in 1988 [1]. Jacobson's idea is to *dynamically* adapt window size to network congestion.

These *end-to-end* delay and loss measurements succinctly summarize the congestion on a path. They are the only local information available for a source to adjust its rate, if the network provides no explicit congestion notification.[2] Moreover, the information is delayed in the sense that the observed delay and loss information at the source reflects the state of the path at an earlier time. Note that the source does not know the delay or loss at individual links in its path. It does not even know (or make use of) its own routing, network topology, or how many other sources are sharing links with it. It must infer congestion from the end-to-end measurements and adjust its rate accordingly.

Similarly, at the links, the local information that is available for the update of congestion measure is the arrival rates of flows that traverse the link. Again, no global information, such as flow rates, delays, or loss probabilities at other links, should be used by a link algorithm. In principle, individual flow rates can be measured and used in the adjustment of congestion measure. However, this would require per-flow processing which can be expensive at high speed. A link algorithm is simpler if it only uses the aggregate rate of all flows traversing the link. We restrict our discussion to this class of link algorithms. For instance, queueing delay and loss probability under first-in-first-out (FIFO) discipline are updated, implicitly, based only on aggregate rate.

[2] Even with ECN bit, RED still provides one-bit of congestion information on the end-to-end path, as DropTail does; see Section 4.2 below. The difference between RED and DropTail is how they generate packet losses, not their information carrying capacity.

4. EXAMPLE: TCP VEGAS

Before we describe mathematical models to understand the equilibrium and stability properties of flow control mechanisms, we briefly describe TCP Vegas, which will be used for illustration later.

4.1. TCP Vegas

Like TCP Reno, TCP Vegas also consists of three phases: slow start, congestion avoidance, and fast retransmit/fast recovery. A Reno source starts cautiously with a small window size of one packet (up to four packets have recently been proposed) and the source increments its window by one every time it receives an acknowledgment. This doubles the window every round-trip time and is called slow start. When the window reaches a threshold, the source enters the congestion avoidance phase, where it increases its window by the reciprocal of the current window size every time it receives an acknowledgment. This increases the window by one in each round-trip time, and is referred to as additive increase. The threshold that determines the transition from slow start to congestion avoidance is meant to indicate the available capacity in the network and is adjusted each time a loss is detected. On detecting a loss through three duplicate acknowledgments, the source sets the slow start threshold to half the current window size, retransmits the lost packet and halves its window size. This is called fast retransmit/fast recover; see Ref. 3 for more details. When the acknowledgment for the retransmitted packet arrives, the source re-enters congestion avoidance. In TCP Reno, slow start is entered only rarely when the source first starts and when a loss is detected by timeout rather than duplicate acknowledgments.

TCP Vegas [5] improves upon TCP Reno through three main techniques. The first is a modified retransmission mechanism where timeout is checked on receiving the first duplicate acknowledgment, rather than waiting for the third duplicate acknowledgment (as Reno would), and results in a more timely detection of loss. The second technique is a more prudent way to grow the window

size during the initial use of slow-start when a connection starts up and it results in fewer losses.

The third technique is a new congestion avoidance mechanism that corrects the oscillatory behavior of Reno. The idea is to have a source estimate the number of its own packets buffered in the path and try to keep this number between α (typically 1) and β (typically 3) by adjusting its window size. The window size is increased or decreased by one in each round-trip time according to whether the current estimate is less than α or greater than β. Otherwise the window size is unchanged. The rationale behind this is to maintain a small number of packets in the pipe to take advantage of extra capacity when it becomes available. Another interpretation of the congestion avoidance algorithm of Vegas is given in Ref. 8, in which a Vegas source periodically measures the round-trip *queueing* delay and sets its rate to be proportional to the ratio of its round-trip propagation delay to queueing delay, the proportionality constant being between α and β. Hence, the more congested its path is, the higher the queueing delay and the lower the rate. The Vegas source obtains queueing delay by monitoring its round-trip time the time between sending a packet and receiving its acknowledgment and subtracting from it the round-trip propagation delay.

4.2. DropTail

Congestion control of the Internet was entirely source-based at the beginning, in that the link algorithm was implicit. A link simply drops a packet that arrives at a full buffer. This is called DropTail (or Tail Drop) and the implicit link algorithm is carried out by the queue process. The congestion measure it updates depends on the TCP algorithm.

For TCP Reno and its variants, the congestion measure is packet loss probability. The end-to-end loss probability is observed at the source and is a measure of congestion on the end-to-end path. For TCP Vegas, the congestion measure turns out to be link queueing delay [8] when FIFO service discipline is used. The congestion measure of a path is the sum of queueing delays at all constituent links.

Random early detection (RED) is proposed in Ref. 9 as an alternative to DropTail. In RED, an arrival packet is discarded with a probability when the average queue length exceeds a minimum threshold, in order to provide early warning of incipient congestion before the buffer overflows. The dropping probability is an increasing function of average queue length. The rationale is that a large average queue length signifies congestion and should intensify the feedback signal. It has also been proposed that a bit in the IP header be used for explicit congestion notification (ECN), so that a link can mark a packet probabilistically (setting the ECN bit from 0 to 1) instead of dropping it.

5. DUALITY MODEL

In this section, we first describe an abstract model for general source and link algorithms. As an illustration; we then applied it to the Vegas/DropTail algorithms described in the last section.

5.1. General Source/Link Algorithms

A network is modeled as a set L of "links," indexed by l, with finite transmission capacities c_l packets per second. It is shared by a set of sources, indexed by s. Each source s is assigned a path along which data is transferred to its destination. A path is a subset of the links and is denoted by $L_s \subseteq L$. For convenience, denote by S_l the subset of sources that traverse link l. Hence, $l \in L_s$ if and only if $s \in S_l$. To understand the equilibrium and stability of the network, we assume for simplicity that the link capacities c_l, the set of links and sources, and the routes L_s are all fixed at the timescale of interest.

Each source s adjusts its transmission rate $x_s(t)$ at time t, in packets per second, based on the congestion on its path. Each link l maintains a measure of congestion $p_l(t)$ at time t. We will call $p_l(t)$ the link *price* for it can be interpreted as unit price for bandwidth at link l (see, [7,10]). A link is said to be congested if $p_l(t)$ has a large value. A path is said to be congested if the sum of link prices is high.

Each source s can observe a delayed version of the sum of link prices, summed over the links in its path. This path price is a measure of congestion in the path end-to-end. Suppose the backward delay from link l to source s is denoted by τ^b_{ls}. Then the path price that is observed by s at time t can be represented by [19]

$$q_s(t) := \sum_{l \in L_s} p_l(t - \tau^b_{ls}) \tag{1}$$

We assume each link l can observe a delayed version of the sum of source rates, summed over the sources that traverse the link. The aggregate rate is a measure of demand for bandwidth at link l. Suppose the forward delay from source s to link l is denoted by τ^f_{ls}. Then the aggregate rate that is observed by link l at time t can be represented by

$$y_l(t) := \sum_{s \in S_l} x_s(t - \tau^f_{ls}) \tag{2}$$

The decentralization requirement dictates that each source s can adjust its rate $x_s(t)$ based only on $q_s(t)$, in addition to its own rate $x_s(t)$. In particular, the rate adjustment cannot depend on individual link prices $p_l(t)$ nor path prices of other sources. This can be modeled as:

$$\dot{x}_s(t) = F_s(x_s(t), q_s(t)) \tag{3}$$

where F_s calculates the amount of rate adjustment. Similarly, each link l can adjust its price $p_l(t)$ based only on $y_l(t)$, in addition to its own price $p_l(t)$. In particular, the price adjustment cannot depend on individual source rates $x_s(t)$ nor aggregate rate at other links. This can be modeled as:

$$\dot{p}_l(t) = G_l(y_l(t), p_l(t)) \tag{4}$$

where G_l represents the (implicit or explicit) price adjustment algorithm at link l. In general, the link and source algorithms can also depend on some internal state variable.

In summary, a general congestion control scheme can be decomposed into two algorithms. The source algorithm that adapts the rate to congestion in its path can be modeled by Eq. (3). The link algorithm that updates the price based on aggregate rate can be modeled by Eq. (4). The information used by these algorithms is not only local, but also delayed as expressed by Eqs. (1) and (2).

5.2. Example: Vegas/DropTail

We now describe a model of Vegas/DropTail, developed in Ref. 8, as an illustration. The model ignores slow start and fast retransmit/fast recovery, and only captures the behavior of congestion avoidance.

The price at link l turns out to represent queueing delay whose dynamics is modeled as

$$\dot{p}_l(t) = \frac{1}{c_l}(y_l(t) - c_l) =: G_l(y_l(t), p_l(t)) \tag{5}$$

at bottleneck links. To model the TCP Vegas algorithm, let d_s be the round-trip propagation delay for source s and assume the Vegas parameters satisfy $\alpha = \beta$ for all sources s. Then the rate is adjusted according to:

$$\dot{x}_s(t) = \frac{1}{(d_s + q_s(t))^2} \operatorname{sgn}\left(1 - \frac{x_s(t)q_s(t)}{\alpha d_s}\right) =: F_s(x_s(t), q_s(t)) \tag{6}$$

where $\operatorname{sgn}(z)$ is -1 if $z < 0$, 0 if $z = 0$, and 1 if $z > 0$. Here, $q_s(t)$ is the (delayed) end-to-end queueing delay in the path of source s, $d_s + q_s(t)$ is the round-trip time observed at source s at time t, and $x_s(t)q_s(t)$ is the number of packets that are buffered in the queues in the path. Hence, Eq. (6) says that the window (rate $\times$ round-trip time) is incremented or decremented at a rate of 1 packet per round-trip time, according as the number $x_s(t)q_s(t)$ of packets buffered in the path is smaller or greater than the target αd_s. In equilibrium, each source s maintains αd_s packets in its path.

6. EQUILIBRIUM AND STABILITY PROPERTIES

Equilibrium properties, such as fairness, utilization and QoS, and dynamic properties, such as stability, of a flow control scheme can be understood by studying the mathematical model specified by Eqs. (1) and (4). In this section, we illustrate how to analyze the model Eqs. (1) and (4).

6.1. Equilibrium

Under mild assumptions on the source algorithm F_s, we can associate a utility function $U_s(x_s)$ with source x_s that is a concave increasing function of its rate x_s. As previously mentioned, this means sources are greedy and there is a diminishing return as rate increases.

Consider the following constrained utility maximization problem:

$$\max_{x \geq 0} \quad \sum_s U_s(x_s) \tag{7}$$

$$\text{subject to} \quad yl \leq c_l \text{ for all links } l \tag{8}$$

The constraint (8) says that the aggregate source rate at any link does not exceed the capacity. From optimization theory, we can associate with the primal problem (7, 8) the following dual problem

$$\min_{p \geq 0} \sum_s \max_{x_s \geq 0}(U_s(x_s) - x_s q_s) + \sum_l c_l p_l \tag{9}$$

where $q_s = \sum_{l \in L_s} p_l$ is the sum of link prices p_l in the path of source s.

Suppose (x^*, p^*) is an equilibrium point of the model (1–4). Then it is proved in [11] that x^* solves the primal problem (Eqs. (7) and (8)) if and only if for all links l

$$y_l^* \leq c_l \quad \text{with equality if } p_l^* > 0 \tag{10}$$

Moreover, in this case, p^* solves the dual problem (9). Note that the condition (10), called *complementary slackness*, says that every bottleneck link is fully utilized, that is, input rate is equalized to capacity.

This interpretation has several implications. First, we can regard the source rates $x_s(t)$ in Eq. (3) as primal variables, the prices $p_l(t)$ in Eq. (4) as dual variables, and a congestion control mechanism (Eqs. (3) and (4)) as a distributed asynchronous computation over a network to solve the primal problem (7–8) and the dual problem (9). The equilibrium rates can be interpreted as utility maximizing, and the equilibrium prices (delay or loss) as Lagrange multipliers that measure the marginal increase in optimal aggregate utility for each unit of increment in link capacities.

Second, different source and link algorithms are just different ways to solve the same prototypical problem (7–9), with different utility functions. Even though TCP algorithms were not designed to solve any optimization problem, they have implicitly chosen certain utility functions by adjusting the source rate in a particular way. Take TCP Vegas algorithm (6) for example: in equilibrium, we have $\dot{x}_s = 0$ and hence $x_s^* q_s^* = \alpha d_s$. This implies a utility function of

$$U_s(x_s) = \alpha d_s \log x_s \tag{11}$$

See Ref. 11 for details and for utility functions of Reno. Moreover, the TCP algorithm F_s alone determines the equilibrium rate allocation by defining the underlying optimization problem. The role of link algorithm G_l is to ensure the complementary slackness condition and to stabilize the equilibrium.

Third, the equilibrium properties are all determined by the underlying optimization problem. Fairness, a property of the optimal rate vector x^*, is determined by the utility functions in the utility maximization (7–8). For Vegas, the log utility function in Eq. (11) implies that it achieves proportional fairness. Hence, as mentioned earlier, we can define fairness through the corresponding utility function. Moreover, since the utility function depends only on source algorithm F_s, fairness is independent of the link algorithm, as long as link prices depend only on aggregate rates $y_l(t)$, not on individual source rates $x_s(t)$.

Fourth, if the link algorithm G_l achieves the complementary slackness condition (10), then the network will be maximally utilized in equilibrium. QoS however may not be properly controlled if prices are coupled with QoS.

In this case, the value of the Lagrange multiplier p_l^* is determined not by the update algorithm G_l, but by the underlying optimization problem. In particular, if the number of sources sharing a link is large, or if the link capacity is small, then p_l^* will be large. If p_l^* represents queueing delay, as in TCP Vegas, or loss probability, as in TCP Reno, QoS can be poor. It is however possible to design link algorithms that decouple congestion measure with QoS such as loss or delay, so that the link prices converge to their equilibrium values determined by the primal problem while queues are kept small; see, Refs. 12 and 13. In this case, price information is no longer embedded in end-to-end delay and must be fed back to sources explicitly.

In summary, equilibrium properties of general source and link algorithms can be understood by regarding them as distributed primal-dual iterations to solve the primal and dual problems (7–9), where the utility functions are determined by the source algorithm F_s.

6.2. Stability

In general, the stability of the distributed nonlinear system with delay, specified by Eqs. (1)–(4), is very difficult to analyze (but see [10]). We can however understand its stability in the presence of delay around an equilibrium by studying the linearized model. In this section, we briefly summarize the stability properties of TCP Reno and TCP Vegas; see, Refs. 14–16 for details.

It is well known that the queue length under TCP Reno can oscillate wildly, with either DropTail or RED link algorithm, and it is extremely hard to reduce the oscillation by tuning RED parameters, Refs. 17 and 18. The additive-increase-multiplicative-decrease (AIMD) strategy employed by TCP Reno (and its variants such as NewReno and SACK) and noise-like traffic that are not effectively controlled by TCP no doubt contribute to this oscillation. It is shown in Ref. 15 however that protocol instability can have much larger effect on the oscillatory behavior than these factors. By instability, we mean severe oscillation in *aggregate* quantities, such as queue length and average window size. The analysis of the linearized delayed model shows that the system becomes unstable when delay increases, and more strikingly, when network capacity increases! This agrees with empirical experience that the current TCP performs poorly at large window sizes. Moreover, even if we smooth out AIMD, that is even if window is not adjusted on each acknowledgment arrival or loss event, but is adjusted periodically by the same *average* amount AIMD would over the same period, the oscillation persists. In particular, this implies that equation-based rate control will not help if the equation mimics the Reno dynamics. This suggests that TCP Reno/RED is ill-suited for future networks where capacities will be large.

This motivates the design of new source and link algorithms that maintain linear stability for general delay and capacity [19–23]. The main insight from this series of work is to scale down source responses with their own round-trip times and scale down link responses with their own capacities, in order to keep the gain over the feedback loop under control.

It turns out that the implicit link algorithm (5) of Vegas has exactly the right scaling with respect to capacity as used in the scalable design of Refs. 19 and 22. This built-in scaling with capacity makes Vegas potentially scalable to high bandwidth, in stark contrast to the AIMD algorithm of Reno and its variants. The source algorithm of Vegas, however, has a different scaling with respect to delay from those in Refs. 19 and 22, making it susceptible to instability in the presence of large delay. It is possible however to stabilize it by slightly modifying the rate adjustment algorithm (6) of Vegas; see Ref. 16 for details.

7. CONCLUSION

Flow control schemes are distributed and asynchronous algorithms to share network resources among competing users. The goal is to share these resources fairly and efficiently, and to provide good QoS in a stable, robust, and scalable manner. There is no automatic method to synthesize flow control schemes that will achieve these objectives. We have provided an introduction to mathematical models that can help understand and design such schemes, and have illustrated these models using TCP Vegas.

Acknowledgments

This article is a gentle introduction to some of the recent literature on flow control. We gratefully acknowledge the contribution of authors of these papers, only some of which are cited here, and in particular, that of my collaborators Sanjeewa Athuraliya, Hyojeong Choe, John Doyle, Ki-baek Kim, David Lapsley, Fernando Paganini, Larry Peterson, Jiantao Wang, Limin Wang, Zhikui Wang. Finally, we acknowledge the support of US National Science Foundation through grant ANI-0113425, US Army Research Office, the Caltech Lee Center for Advanced Networking, and Ciso.

BIOGRAPHY

Steven. H. Low received his B.S. degree from Cornell University and PhD from the University of California–Berkeley, both in electrical engineering. He was with AT&T Bell Laboratories, Murray Hill, from 1992 to 1996, with the University of Melbourne, Australia, from 1996 to 2000, and is now an associate professor at the California Institute of Technology, Pasadena. He was a corecipient of the IEEE William R. Bennett Prize Paper Award in 1997 and the 1996 R&D 100 Award. He is on the editorial board of IEEE/ACM Transactions on Networking. He has been a guest editor of the IEEE Journal on Selected Area in Communications, on the program committee of major networking conferences. His research interests are in the control and optimization of communications networks and protocols. His home is netlab.caltech.edu and email is slow@caltech.edu.

BIBLIOGRAPHY

1. V. Jacobson, Congestion avoidance and control, *Proc. SIGCOMM'88, ACM*, August 1988. An updated version is available via *ftp://ftp.ee.lbl.gov/papers/congavoid.ps.Z*.
2. E. Varvarigos and T. Varvarigou, Computer communications protocols, in J. G. Proakis, ed., New York, *Encyclopedia of Telecommunications*, Wiley, 2002.

3. J. Aweya, Transmission control protocol, in John Proakis, ed. *Encyclopedia of Telecommunications*, New York, Wiley, 2002.

4. E. J. Hernandez-Valencia, L. Benmohamed, R. Nagarajan, and S. Chong, Rate control algorithms for the ATM ABR service, *Eur. Trans. Telecomm.* **8**: 7–20 (1997).

5. L. S. Brakmo and L. L. Peterson. TCP Vegas: end-to-end congestion avoidance on a global Internet, *IEEE J. Select. Areas Comm.* **13**(8): 1465–1480 (October 1995) *http://cs.princeton.edu/nsg/papers/jsac-vegas.ps.*

6. D. Bertsekas and R. Gallager. *Data Networks*, 2nd ed. Prentice-Hall, 1992.

7. F. P. Kelly, A. Maulloo, and D. Tan, Rate control for communication networks: Shadow prices, proportional fairness and stability, *J. Operations Res. Soc.* **49**(3): 237–252 (March 1998).

8. S. H. Low, L. L Peterson, and L. Wang, Understanding Vegas: a duality model, *J. ACM* **49**(2): 207–235 (March 2002). *http://netlab.caltech.edu.*

9. S. Floyd and V. Jacobson, Random early detection gateways for congestion avoidance, *IEEE/ACM Trans. Networking* **1**(4): 397–413 (August 1993). *ftp://ftp.ee.lbl.gov/papers/early.ps.gz.*

10. S. H. Low and D. E. Lapsley, Optimization flow control, I: basic algorithm and convergence, *IEEE/ACM Trans. Networking* **7**(6): 861–874, (December 1999). *http://netlab.caltech.edu.*

11. S. H. Low, A duality model of TCP and queue management algorithms, In *Proc. ITC Specialist Seminar on IP Traffic Measurement, Modeling and Management* (*updated version*) (September 18–20, 2000). *http://netlab.caltech.edu.*

12. S. Athuraliya, V. H. Li, S. H. Low, and Q. Yin, REM: active queue management. *IEEE Network* **15**(3): 48–53 (May/June 2001). Extended version in *Proc. ITC17*, Salvador, Brazil, September 2001. *http://netlab.caltech.edu.*

13. C. Hollot, V. Misra, D. Towsley, and W. B. Gong, On designing improved controllers for AQM routers supporting TCP flows. In *Proc. IEEE Infocom* (April 2001). *http://www-net.cs.umass.edu/papers/papers.html*.

14. C. Hollot, V. Misra, D. Towsley, and W. B. Gong, A control theoretic analysis of RED. In *Proc. of IEEE Infocom* (April 2001). *http://www-net.cs.umass.edu/papers/papers.html*.

15. S. H. Low et al., Dynamics of TCP/RED and a scalable control. In *Proc. IEEE Infocom* (June 2002). *http://netlab.caltech.edu.*

16. H. Choe and S. H. Low. *Stabilized Vegas*, In *Proc. of 39th Annual Allerton Conference on Communication, Control, and Computing*, (October 2002). *http://netlab.caltech.edu.*

17. M. May, T. Bonald, and J.-C. Bolot, Analytic evaluation of RED performance, In *Proc. IEEE Infocom* (March 2000).

18. M. Christiansen, K. Jeffay, D. Ott, and F. D. Smith, Tuning RED for web traffic, In *Proc. ACM Sigcomm* (2000).

19. F. Paganini, John C. Doyle, and S. H. Low, *Scalable laws for stable network congestion control*. In *Proc. Conference on Decision and Control* (December 2001). *http://www.ee.ucla.edu/paganini.*

20. G. Vinnicombe, On the stability of end-to-end congestion control for the Internet. Technical report, Cambridge University, CUED/F-INFENG/TR.398, (December 2000).

21. G. Vinnicombe, Robust congestion control for the Internet. Submitted for publication, 2002.

22. F. Paganini, Z. Wang, S. H. Low, and J. C. Doyle. A new TCP/AQM for stability and performance in fast networks. In *Proc. of 39th Annual Allerton Conference on Communication, Control, and Computing* (October 2002).

23. S. Kunniyur and R. Srikant. A time-scale decomposition approach to adaptive ECN marking. *IEEE Trans. Automatic Control* (June 2002).

NETWORK RELIABILITY AND FAULT TOLERANCE

MURIEL MÉDARD
Massachusetts Institute of Technology
Cambridge, Massachusetts

STEVEN S. LUMETTA
University of Illinois
Urbana — Champaign
Urbana, Illinois

1. INTRODUCTION

The majority of communications applications, from cellular telephone conversations to credit card transactions, assume the availability of a reliable network. At this level, data are expected to traverse the network and to arrive intact at their destination. The physical systems that compose a network, on the other hand, are subjected to a wide range of problems, ranging from signal distortion to component failures. Similarly, the software that supports the high-level semantic interface often contains unknown bugs and other latent reliability problems. Redundancy underlies all approaches to fault tolerance. Definitive definitions for all concepts and terms related to reliability, and, more broadly, dependability, can be found in the book by Anderson et al. [1].

Designing any system to tolerate faults first requires the selection of a fault model, a set of possible failure scenarios along with an understanding of the frequency, duration, and impact of each scenario. A simple fault model merely lists the set of faults to be considered; the decision regarding inclusion in the set is based on a combination of expected frequency, impact on the system, and feasibility or cost of providing protection. Most reliable network designs address the failure of any single component, and some designs tolerate multiple failures. In contrast, few attempt to handle the adversarial conditions that might occur in a terrorist attack, and cataclysmic events are almost never addressed at any scale larger than a city.

The temporal characteristics of faults vary widely, but can be roughly categorized as permanent, intermittent, or transient. Failures that prevent a component from functioning until repaired or replaced, such as the destruction of a network fiber by a backhoe, are considered permanent. Failures that allow a component to function properly some of the time are called *intermittent*. Damaged connectors and electrical components sometimes produce intermittent faults, operating correctly until mechanical vibrations or thermal variations cause a failure, and recovering when conditions change again. The last category, transient

faults, is usually the easiest to handle. Transient faults range from changes in the contents of computer memory due to cosmic rays, or bit errors due to thermal noise in a demodulator, and are typically infrequent and unpredictable. The difference between an intermittent fault and a transient fault is sometimes solely one of frequency; for transient faults, a combination of error-correcting codes and data retransmission usually provides adequate protection.

Redundancy takes two forms, spatial and temporal. *Spatial redundancy* replicates the components or data in a system. Transmission over multiple paths through a network and the use of error correction codes are examples of spatial redundancy. Temporal redundancy underlies automatic repeat request (ARQ) algorithms, such as the sliding-window abstraction used to support reliable transmission in the Internet's Transmission Control Protocol (TCP). A reliable network typically provides both spatial and temporal redundancy to tolerate faults with differing temporal persistence. Spatial redundancy is necessary to overcome permanent failures in physical components, while temporal redundancy requires fewer resources and is thus preferable when dealing with transient errors.

Beyond the selection of a fault model, several additional problems must be considered in the design of a fault-tolerant system. A system must be capable of detecting each fault in the model, and must be able to isolate each fault from the functioning portion of the system in a manner that prevents faulty behavior from spreading. As a fault detection mechanism may detect more than one possible fault, a system must also address the process of fault diagnosis (or localization), which narrows the set of possible faults and allows more efficient fault isolation techniques to be employed. An error identified by a system need not necessarily be narrowed down to a single possible fault, but a smaller set of possibilities usually allows a more efficient strategy for recovery.

Fault isolation boundaries are usually designed to provide fail-stop behavior for the desired fault model. The term *fail stop* implies that incorrect behavior does not propagate across the fault isolation boundary; instead, failed components cease to produce any signals. Fail stop does not imply self-diagnosis; components adjacent to a failed component may diagnose the failure and deliberately ignore any signals from the failed component, but the physical system design must allow such a decision. In a router, for example, the interconnect between cards controlling individual links must provide electrical isolation to support fail-stop behavior for failed cards. A bus-based computer interconnect does not allow for fail stop, as nothing can prevent a failed card from driving the bus lines inappropriately. In modern, high-end servers, such buses have been replaced by switched networks with broadcast capability in order to enable such isolation. The eradication of similar phenomena in the move from shared to switched Ethernets in the mid-1990s was one of the main administrative advantages of the change, as failed hosts are much less likely to render a switched network unusable by flooding it with continuous traffic.

Two models of network service have dominated research and commercial networking. The first is the telephony network, or more generally a network in which quasipermanent routes called *circuits* deliver fixed data capacity from one point to another. In digital telephony, a voice circuit requires 64 kbps (kilobits per second); a single lightpath in a wavelength-division-multiplexed (WDM) optical network may deliver up to 40 Gbps, but is conceptually similar to the circuit used to carry a phonecall. The second network service model is the packet-switched data network, which evolved from the early ARPANET and NSFNET projects into the modern Internet. Packet-switched networks seldom provide strong guarantees on delivered data rate or maximum delay, but are typically more efficient than circuit-oriented designs, which must base guaranteed agreements on worst-case traffic load scenarios.

For the purposes of our discussion, the key difference between these two models lies in the fact that applications using packet-switched networks can generally tolerate more serious service disruptions than can those based on circuit-switched networks. The latter class of applications may assume that data rate, delay, and jitter guarantees provided by the network will be honored even when failures occur, whereas minor disruptions may occur even in normal circumstances on packet-switched networks because of fluctuations in traffic patterns and loads. Fault tolerance issues are thus addressed in markedly different ways in the two types of networks. In packet-switched networks like the Internet, users currently tolerate restoration times of minutes [2,3], whereas fault tolerance for circuit-switched networks can be considered a component of quality of service (QoS) [4,5], and is typically achieved in milliseconds, or, at worst, seconds.

The majority of this article focuses on fault tolerance issues in high-speed backbone networks, such as wide-area networks (WANs) and metropolitan-area networks (MANs). Such networks are predominantly circuit-based and carry heavy traffic loads. As even a short downtime may cause substantial data loss, rapid recovery from failure is important, and these networks require high levels of reliability. Backbone networks generally are implemented using optical transmission and, conversely, fault tolerance in optical networks is typically considered in the context of backbone networks [6,7]. In these networks, a failure may arise because a communications link is disconnected or a network node becomes incapacitated. Failures may occur in military networks under attack [8], as well as in public networks, in which failures, albeit rare, can be extremely disruptive [9, Chap. 8].

The next section provides an overview of fault detection mechanisms and the basic strategies available for recovery from network component failures. Sections 3 and 4 build on these basics to illustrate recovery schemes for high-speed backbone networks. Sections 5 and 6 examine simple and more complex topologies and discuss the relationship between topology and recovery. Section 5 highlights ring topologies, as they are a key architectural component of high-speed networks. Section 6 extends the concepts developed for rings by overlaying logical ring topologies over physical mesh topologies. We also discuss

some link- and node-based reliability schemes that are specifically tailored to mesh networks. Although the text focuses on approaches to fault tolerance in high-speed backbone networks, many of the principles also apply to other types of networks. In Section 7, we move away from circuit-switched networks and examine fault tolerance for packet-switched networks, and in particular the Internet. Finally, Section 8 discusses reliability issues for local-area networks (LANs).

2. FAILURE DETECTION AND RECOVERY

A wide variety of approaches have been employed for detection of network failures. In electronic networks with binary voltage encodings (e.g., RS-232), two nonzero voltages are chosen for signaling. A voltage of zero thus implies a dead line or terminal. Similarly, electronic networks based on carrier modulation infer failures from the absence of a carrier. Shared segments such as Ethernet have been more problematic, as individual nodes cannot be expected to drive the segment continuously. In such networks, many failures must be detected by higher levels in the protocol stack, as discussed later in this section.

The capacity of optical links makes physical monitoring a particularly important problem, and many techniques have been explored and used in practice. Optical encoding schemes generally rely on on/off keying; that is, the presence of light provides one signal, and its absence provides a second. With single-wavelength optics, information must be incorporated into the channel itself. One approach is to monitor time-averaged signal power, using an encoding scheme that results in a predictable distribution of ON and OFF frequencies. A second approach utilizes overhead bits in the channel, allowing bit error rate (BER) sampling at the expense of restricting the data format used by higher levels of the protocol stack. A third approach employs a sideband to carry a pilot tone. These approaches are complementary, and can be used in tandem.

A WDM system typically applies the single-wavelength techniques just mentioned to each wavelength, but the possibility of exploiting the multiplexing to reduce the cost of failure detection has given rise to new techniques. A single wavelength, for example, can be allocated to provide accurate estimates of BER along a link. Unfortunately, this approach may fail to detect frequency-dependent signal degradation. Pairing of monitoring wavelengths with data wavelengths reduces the likelihood of missing a frequency-dependent failure, but is too inefficient for most networks.

The approaches discussed so far have dealt with failure detection at the link level. With circuit-switched networks, the receiver on any given path can directly monitor accumulated effects along the entire path. The techniques discussed for a single wavelength can also be employed for a full path with optically transparent networks. With networks that perform optoelectronic conversion at each node, only in-band information is retained along the length of the path and overhead in the data format is typically necessary for failure detection. Path-based approaches are advantageous in the sense that they may cover a broader set of possible failures. They get to the root of the problem; something went wrong getting from the sender to the receiver. Link-based approaches, however, make fault localization simpler, an important benefit in finding and repairing problems in the network. In practice, most backbone networks use a combination of link and path detection techniques to obtain both benefits.

Additional fault tolerance is often included in higher levels of a network protocol stack. Most protocols used for data networking (as opposed to telephony), for example, include some redundancy coding for the purposes of error detection. Typically, feedback from these layers is not provided to the physical layer, although some exceptions do exist in LANs, such as the use of periodic packet transmissions and inference of failures when no packet arrives (see Section 8 for more detail). Instead, the error detection schemes allow the network to tolerate transient errors through temporal redundancy, namely, retransmission. Voice channels and other redundant forms of data also utilize error correction or other error tolerance techniques in some cases. A telephone circuit crossing an asynchronous transfer mode (ATM) network may lose an occasional cell to a cyclic redundancy check (CRC) failure. In such a case, the cell is discarded, and the voice signal regenerated by interpolation from adjacent cells. This interpolation suffices to make a single cell loss undetectable to humans; thus, as long as the transient errors occur infrequently, no loss is noticed by the people using the circuit.

The choice of failure detection methods used in a backbone network is intertwined with the choice of strategies for restoring circuits that pass through a failed element of the network. Path monitoring, for example, does not readily provide information for failure localization. Correlated failures between paths may help to localize failures, but typically a more careful investigation must be initiated to find the problem. Path monitoring also requires that failure information propagate to the endpoints of the path, delaying detection. Link monitoring allows more rapid and local response to failures, but does not require such an approach. Instead, failure information can be propagated to the ends of each path crossing a link, while the localized failure information is retained for initiating repairs and for dynamic construction of future paths. At the algorithmic level, circuit rerouting schemes can be broadly split into path-based and link- or node-based approaches.

Prompted by the increasing reliance on high-speed communications and the requirement that these communications be robust to failure, backbone networks have generally adopted self-healing strategies to automatically restore functionality. The study of self-healing networks is often classified according to the following three criteria [e.g., 10,11]: (1) the use of link (line) rerouting versus path (or end-to-end) rerouting, (2) the use of centralized computation versus distributed computation, and (3) the use of precomputed versus dynamically computed routes. A succinct comparison of the different options can be found in the book by Wu [12, pp. 291–294] and the paper by Johnson et al. [13]. For path recovery, when a failure leaves a node disconnected from the primary route, a backup route, which may or may not share nodes and links with the primary route, is used. *Link rerouting* usually refers

to the replacement of a link by links connecting the two end nodes of the failed link. When the rerouting is precomputed, the method is generally termed *protection*. Thus, *path protection* refers to precomputed recovery applied to connections following a particular path across a network. *Link or node protection* refers to precomputed recovery of all the traffic across a failed link or node, respectively. Figure 1 illustrates path and link rerouting. Protection routes are precomputed at a single location, and are thus centralized, although some distributed reconfiguration of optical switches may be necessary before traffic is restored. Restoration techniques, on the other hand, can rely on distributed signaling between nodes or on allocation of a new path by a central manager.

3. PATH-BASED SCHEMES

Protection schemes, in which recovery routes are preplanned, generally offer better recovery speeds than restoration approaches, which search for new routes dynamically in response to a failure and generally involve software processing [14,15]. The *Synchronous Optical* NET*work* (SONET) specification, for example, requires that recovery time with protection approaches be under 60 ms. Recovery can be achieved in tens of milliseconds using optomechanical add/drop multiplexers [16,17], and in a few microseconds using acoustooptical switches [18,19]. In contrast, dynamic distributed restoration using digital cross-connect systems (DCSs) for ATM or SONET [20–23] typically targets a 2-s recovery-time goal [17,24,25]. Dynamic centralized path restoration for SONET [26] may even take minutes [24,27]. The performance of several algorithms has been reviewed [28,29]. Restoration typically requires less protection capacity, however.

In this section, we focus on path protection, as the majority of current backbone networks utilize such techniques. Path protection trades longer recovery times for reduced capacity requirements relative to the link-based approaches discussed in the next section. These tradeoffs are discussed in more depth elsewhere [30–32]. Path protection involves finding, for each circuit, a backup route (or path). Figure 2 shows two primary routes and their corresponding backup routes. For each circuit, the two routes do not overlap on any links, implying that no single link failure can affect both a primary route and its backup.

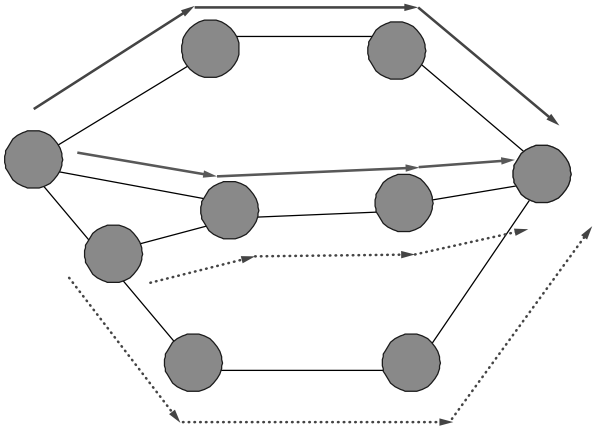


Figure 2. Path protection and associated bandwidth sharing on the backup.

Path protection can itself be divided into several categories: one-plus-one (written $1+1$), one-for-one (written $1:1$), and one-for-N ($1:N$). In the first case, all data are sent along two routes simultaneously, a primary and a backup. The two routes are link-disjoint for tolerance to link failures, or node-disjoint (excluding source and destination) for tolerance to node failures. The receiver monitors incoming traffic on the primary route; when a component along the primary route fails, the receiver switches to the backup signal. The backup route is typically the longer of the two, ensuring that no data are lost because of a single failure. Because both primary and backup routes carry live traffic, the $1+1$ approach is sometimes referred to as *live backup*. Recovery using live backup is extremely fast, as it is based on a local decision by a single node. Protection capacity requirements are high, however, as the backup channel cannot be shared among connections.

The other two approaches, together known as *event-triggered backup*, require less network capacity devoted to protection than does live backup. The penalty is loss of some data when a failure occurs as well as slower restoration times relative to live backup. With event-triggered backup, the backup path is activated only after a failure is detected. As with live backup, the receiver monitors the primary path, but rather than acting locally when it detects a failure, the receiver notifies the sender that a failure has occurred on the primary path, at

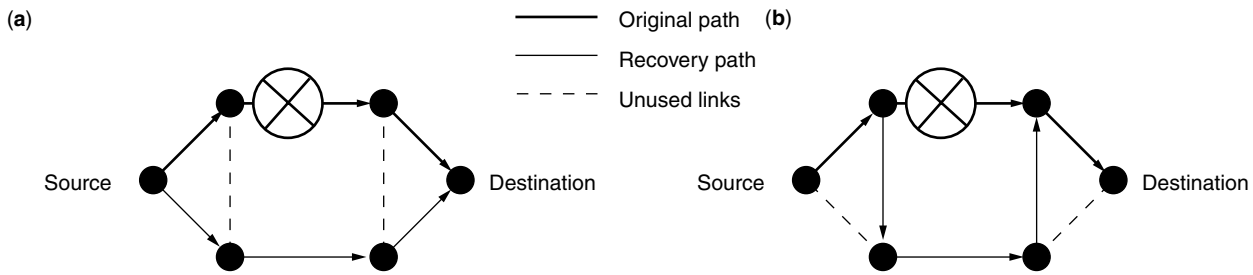


Figure 1. Path (**a**) and link (**b**) rerouting; the failure is marked with an ⊗.

which point the sender begins sending traffic on the backup path. All data transmitted or in flight between the time of the failure and the sender switching over to the backup route are lost. With $1:1$ protection, optical cross-connects are preconfigured for a particular route. Sharing and reuse of the backup route is therefore somewhat restricted. With $1:N$ protection, backup resources can be shared between any set of circuits for which the primary routes do not have resources in common, as illustrated by the two circuits in Fig. 2. Two primary routes passing through a single link, for example, cannot share backup resources; if that link should fail, only one of the two routes could be recovered. The need to configure optical cross-connects along the backup route adds further delay to restoration time and increases data loss with $1:N$ protection, but sharing of backup resources reduces protection capacity requirements by roughly 15–30% relative to $1+1$ protection in an all-optical network. Traffic grooming, in which traffic can be assigned to wavelengths in granularities smaller than those of whole wavelengths, can allow for much more effective sharing.

All forms of protection require adequate spatial redundancy in the network topology to allow advance selection of two disjoint routes for each circuit. For link (or node) protection, this requirement translates to a need for two-edge (-vertex)-redundant graphs; in other words, a graph must remain completely connected after removal of any single edge (vertex, along with all adjacent edges). For path protection, the same condition suffices, as shown by Menger's theorem [33,34], which states that, given a two-edge (-vertex)-redundant graph, two edge- (-vertex)-disjoint routes can be found between any two vertices in the graph. A variety of schemes based on Menger's theorem have been proposed, such as subnetwork connection protection (SNCP) and variations thereof [35–40], which establish two paths between every pair of nodes. However, Menger's theorem is only the starting point for designing a recovery algorithm, which must also consider rules for routing and reserving spare capacity. Path protection over arbitrary redundant networks can also be performed with trees, for example, which are more bandwidth-efficient for multicast traffic [41,42].

With ATM, path rerouting performed by the private network node interface (PNNI) tears down virtual circuit (VC) connections after a failure, forcing the source node to establish a new end-to-end route. Backup virtual paths (VPs) can be predetermined [43] or selected jointly by the end nodes [44]. Source routing, which is used by ATM PNNI, can be preplanned [45] or partially preplanned [46].

4. LINK- AND NODE-BASED SCHEMES

As with path rerouting, methods commonly employed for link and node rerouting in high-speed networks can be divided into protection and restoration, although some hybrid schemes do exist [23]. The two types offer a tradeoff between adaptive use of backup (or "spare") capacity and speed of restoration [25,46]. Dynamic restoration typically involves a search for a free path using backup capacity [47–49] through broadcasting of help messages [13,20,22,24,25,28]. The performance of several algorithms has also been discussed [28,29]. Overheads due to message passing and software processing render dynamic processing slow. For dynamic link restoration using digital cross-connect systems, a 2-s restoration time is a common goal for SONET [17,20–22,24,25]. Preplanned methods, or link protection, depend mostly on lookup tables and switches or add/drop multiplexers. For all-optical networks, switches may operate in a matter of microseconds or nanoseconds and propagation delay dominates switching time.

Link and node protection can be viewed as a compromise between live and event-triggered path protection. Although not as capacity-efficient as $1:N$ path protection [32,44], link protection is more efficient than live path backup, as backup capacity is shared between links. All traffic carried by a failed link or node is recovered independent of the circuits or end-to-end routes associated with the traffic. In particular, the two nodes adjacent to the failure initiate recovery, and only nodes local to the failure typically take part in the process. Backup is not live, but triggered by a failure. Overviews of the different types of protection and restoration methods and comparison of the tradeoffs among them can be found elsewhere in the literature [15,30,50–52].

The fact that link and node protection are performed independently of the particular traffic being carried does provide an additional benefit. In particular, these approaches are independent of traffic patterns, and can be preplanned once to support arbitrary dynamic traffic loads. Path protection does not provide this feature; new protection capacity may be necessary to support additional circuits, and routes chosen without knowledge of the entire traffic load, as is necessary when allocating routes online, are often suboptimal. This benefit makes link and node restoration particularly attractive at lower layers, at which network management at any given point in the network may not be aware of the origination and destination, or of the format [24] of all the traffic being carried at that location.

Link rerouting in ATM usually involves a choice of new routes by nodes adjacent to the failure [53,54].

5. RINGS

Rings have emerged as one of the most important architectural building blocks for backbone networks in the MAN and WAN arenas. While ring networks can support both path-based and link- or node-based schemes for reliability, rings merit a separate discussion because of the practical importance and special properties of ring architecture.

Rings are the most common means of implementing both path and link protection in SONET, which is the dominant protocol in backbone networks. The building blocks of SONET networks are generally self-healing rings (SHRs) and diversity protection (DP) [12,16,19,24,55–64]. SHRs are unidirectional path-switched rings (UPSRs) or bidirectional line-switched rings (BLSRs), while DP refers to physical redundancy in which a spare link (node) is assigned to one or several links (nodes) [12, pp. 315–322].

In SONET, path protection is usually performed by a UPSR, as illustrated by Fig. 3a, in which a route in the

clockwise direction is replaced by a route in the counterclockwise direction in response to a failure. Link rerouting is performed in a SONET BLSR using a technique known as *loopback*, in which traffic is sent back in the direction from which it came. Figure 3b illustrates this operation, in which a route in the clockwise direction is looped back (redirected) onto a counterclockwise route at the node adjacent to the failure. After traveling around the entire ring to the other end of the failed link, the route is looped back again away from the failed link, rejoining the original route. Note that, with the exception of the failed link, the final backup route includes all of the original route. The waste of bandwidth due to traversing both directions on a single link, known as *backhauling*, can be eliminated by looping back at points other than the failure location [65,66]. In case of failure of a node on a BLSR, the failure is handled in a manner similar to that for a link failure. The failure of a node is equivalent to the failure of a metalink consisting of the node and the two links adjacent to it. The only difference is that network management must be able to purge any traffic directed to the failed node from the network.

Loopback operations can be performed on entire fibers or on individual wavelengths. With fiber-based loopback, all traffic carried by a fiber is backed by another fiber, regardless of how many wavelengths are involved. If traffic is allowed in both directions in a network, fiber-based loopback relies on four fibers, as illustrated in Fig. 4. In WDM-based loopback, restoration is performed in a wavelength-by-wavelength basis.

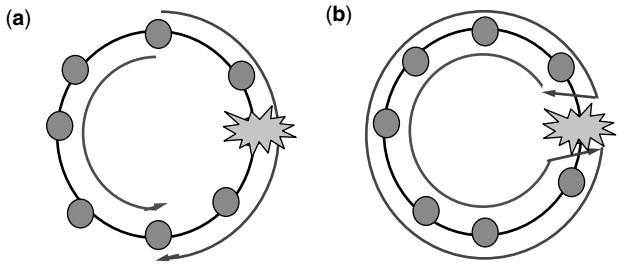


Figure 3. Path protection and node protection in a ring: (**a**) UPSR — automatic path switching; (**b**) BLSR — link/node rerouting.

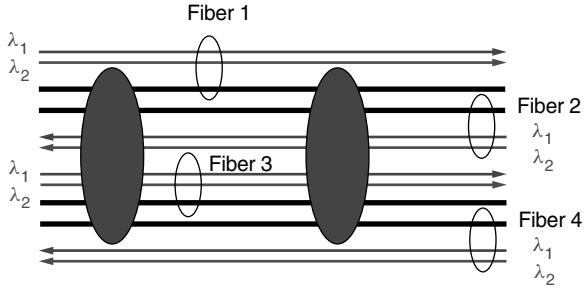


Figure 4. Four-fiber system with fiber-based loopback. Primary traffic is carried by fiber 1 and fiber 2. Backup is provided by fiber 3 for fiber 1 and by fiber 4 for fiber 2.

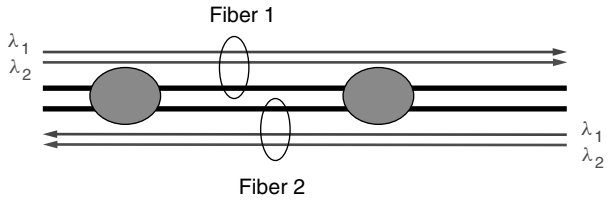


Figure 5. Two-fiber WDM-based loopback. Primary traffic is carried by fiber 1 on λ_1 and by fiber 2 on λ_2. Backup is provided by on λ_1 on fiber 2 for λ_1 on fiber 1. λ_2 on fiber 2 is backed up by λ_2 on fiber 1.

WDM-based loopback requires at least two fibers. Figure 5 illustrates WDM-based loopback. A two-fiber counterpropagating WDM system can be used for WDM-based loopback, even if traffic is allowed in both directions. Note that WDM loopback as shown in Fig. 5 does not require any change of wavelength; traffic initially carried by λ_1 is backed up by the same wavelength. Obviating the need for wavelength changing is economical and efficient in WDM networks. One could, of course, back up traffic from λ_1 on fiber 1 onto λ_2 on fiber 2, if there were advantages to such wavelength changing, for instance in terms of wavelength assignment for certain traffic patterns. We can easily extend the model to a system with more fibers, as long as the backup for a certain wavelength on a certain fiber is provided by some wavelength on another fiber. Moreover, we may change the fiber and/or wavelengths from one fiber section to another. For instance, the backup to λ_1 on fiber 1 may be λ_1 on fiber 2 on a two-fiber section and λ_2 on fiber 3 on another section with four fibers. Note, also, that we could elect not to back up λ_1 on fiber 1 and instead use λ_1 on fiber 1 for primary traffic. The extension to systems with more fibers, interwavelength backups and backups among fiber sections can be readily done.

The finer granularity of WDM-based recovery systems provides several advantages over fiber-based systems. First, if fibers carry at most half of their total capacity, only two fibers rather than four are necessary to provide recovery. Thus, a user need only lease two fibers, rather than paying for unused bandwidth over four fibers. On existing four-fiber systems, fibers could be leased by pairs rather than fours, allowing two leases of two fibers each for a single four-fiber system. The second advantage is that, in fiber based-systems, certain wavelengths may be selectively given restoration capability. For instance, half the wavelengths on a fiber may be assigned protection, while the rest may have no protection. Different wavelengths may thus afford different levels of restoration QoS, which can be reflected in pricing. In fiber-based restoration, all the traffic carried by a fiber is restored via another fiber. If each fiber is less than half full, WDM-based loopback can help avoid the use of counterpropagating wavelengths on the same fiber. Counterpropagating wavelengths on the same fiber are intended to enable duplex operation in situations that require a full fiber's worth of capacity in each direction and that have scarce fiber resources. However, counterpropagation on the same fiber is onerous and reduces the number of wavelengths that a fiber can carry with respect to unidirectional propagation. WDM-based loopback may make using two unidirectional

fibers preferable to using two counterpropagating fibers, for which one fiber is a backup for the other.

When more than one ring is required, rings must be interconnected. In SONET, the usual method to handle nodes shared between rings is called *matched nodes*. Figure 6 shows matched nodes under normal operating conditions. Consider traffic moving from ring 1 to ring 2; traffic in the reverse direction is handled similarly. Under normal operation, matched node 1 is responsible for all interring communications. Matched node 1 houses an add/drop multiplexer (ADM) that performs a drop and continue operation. The drop and continue operation consists of duplicating all traffic through matched node 1 and transmitting it to matched node 2. Thus, matched node 2 has a live backup of all the traffic arriving to matched node 1, and mirrors the operation of matched node 1. However, under normal operating conditions, ring 2 disregards the output from matched node 2. Failure of any node other than the primary matched node is handled by a single ring in a standalone manner. Failure of the secondary matched node treats intraring and interring traffic differently. Note that, depending on the failure mode of the primary matched node, the failure may be seen by both rings or by a single ring. Indeed, failures may occur only on access cards interfacing with one or the other ring, or a wholesale failure may be detected by both rings. Loopback is performed on all the rings that see the failure. Intraring traffic is recovered within the ring wherein it lies, and interring traffic is handled by the second node. How to extend matched nodes to cases other than simple extensions of the topology shown in Fig. 6 is generally unknown. For instance, the case in which a node is shared by more than one ring is difficult. Similarly, the case in which two adjacent rings share links without duplication of resources, as shown in Fig. 8, is complicated in the case of shared nodes.

Many schemes besides SONET exist or have been proposed to enable ring-based networks, usually using optical fiber as the medium. The proprietary protocol Fiber Distributed Data Interface (FDDI) [67–69] is such a scheme. Both FDDI and IEEE 802.5 control access to the ring by passing an electronic token from node to node. Only the node with the token is allowed to transmit. Multiple ring topologies may be interconnected through a hub [70], or rings may coexist in a logically interconnected fashion over a single physical ring [71–73], or rings may be arranged hierarchically [70,74,75].

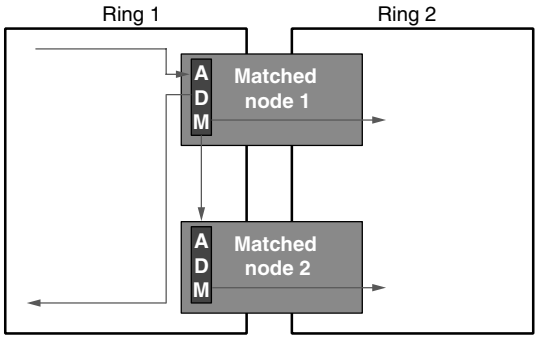


Figure 6. Matched nodes.

The IEEE 802.17 Resilient Packet Ring (RPR) Working Group has been set up to investigate the use, mainly at the MAN, of an optical ring architecture coupled with a packet-based MAC. The purpose of this project is to combine the robustness of rings with a flexible MAC that is well suited to current optical access applications [76].

6. MESH NETWORKS

In this section, we expand our exploration of topologies to redundant meshes. Restricting a network to use only DP and SHRs is a constraint that has cost implications for building and expanding networks [63]; see Stoer's book [34] for an overview of the design of topologies under certain reliability constraints. Ring-based architectures may be more expensive than meshes [63,77], and as nodes are added, or networks are interconnected, ring-based structures may be difficult to preserve, thus limiting their scalability [12,63,78]. However, rings are not necessary to construct fault-tolerant networks [79,80]. Mesh-based topologies can also provide redundancy [34,78,81]. Even if we constrain ourselves to always use ring-based architectures, such architectures may not easily bear changes and additions as the network grows. For instance, adding a new node, connected to its two nearest node neighbors, will preserve mesh structure, but may not preserve ring structure. Our arguments indicate that, for reasons of cost and extensibility, mesh-based architectures are more promising than interconnected rings.

Algorithmic approaches to general mesh restoration are often difficult, however, and implementations can be substantially more complex. To address this problem, many techniques attempt to find rings within the meshes. Overlays using rings are obtained by placing cycles atop existing mesh networks. Each such cycle creates a ring. Service protection or restoration is then generally obtained on each ring as though it were a physical ring. Covering mesh topologies with rings is a means of providing both mesh topologies and distributed, ring-based restoration. Numerous approaches ensure link restorability by finding covers of rings for networks. Many of these techniques have been proposed in the context of backbone networks in order to enable recovery over mesh topologies.

One such approach is to cover nodes in the network by rings [56]. In this manner, a portion of links are covered by rings. If primary routings are restricted to the covered links, link restoration can be effected on each ring in the same manner as in a traditional SHR, by routing backup traffic around the ring in the opposite direction to the primary traffic. Using such an approach, the uncovered links can be used to carry unprotected traffic; that is, traffic that may not be restored if the link that carries it fails. However, under some conditions it may not be possible to cover all nodes with a single ring, or the length of the resulting may be undesirable. A large ring forces long routes for many connections. Such long routes have several drawbacks, both from the point of view of routing (reduced wavelength-assignment efficiency) and from the point of view of communications (excessive jitter).

To allow every link to carry protected traffic, other ring-based approaches ensure that every link is covered by a

ring. One approach to selecting such covers is to cover a network with rings so that every link is part of at least one ring [82]. Several issues arise concerning the overlap and interconnection of such rings. Many of these issues are similar to issues encountered in SONET network deployment. The two main issues are management of links logically included in two rings and node management for ring interconnection.

The first issue concerns the case in which a single link is located on two rings. If that link bears a sufficient number of fibers or wavelengths, the two rings can be operated independently over that link, as shown in Fig. 7. However, the resources available to the overlay network may require sharing the resources over that link. Figure 8 shows such a network, in which only a single wavelength is available to the overlay network. In such a case, the logical fibers must be physically routed through available physical fibers, with network management acting to ensure that conflicts are avoided on the shared span. Such operations incur significant overhead.

The second issue relates to node interconnection among rings. Minimizing the amount of fiber required to obtain redundancy using ring covers is equivalent to finding the minimum cycle cover of a graph, an NP-complete problem [83,84], although bounds on the total length of the cycle cover may be found [85].

A more recent approach to ring covers, intended to overcome the difficulties of previous approaches, is to cover every link with exactly two rings, each with two fibers. The ability to perform loopback style restoration over any link in mesh topologies was first introduced in the late 1990s [86,87], using certain types of ring covers. In particular, Ellinas et al. [86] consider link failure restoration in optical networks with arbitrary two-link redundant arbitrary mesh topologies and bidirectional links. The approach is an application of the double-cycle ring cover [88–90], which selects cycles in such a way that each edge is covered by two cycles. For planar graphs, the problem can be solved in polynomial time; for nonplanar graphs, it is conjectured that double-cycle covers exist, and a counterexample must have certain properties unlikely to occur in network graphs [91]. Cycles can be used as rings to perform restoration. Each cycle corresponds to either a primary or a secondary two-fiber ring. Let us consider a link covered by two rings, rings 1 and 2. If we assign a direction to ring 1 and the opposite direction to ring 2, ring-based recovery using the double-cycle cover uses ring 2 to back up ring 1. This recovery technique is similar to recovery in conventional SHRs, except that the two rings that form four-fiber SHRs are no longer collocated over their entire lengths. In the case of four fiber systems, with two fibers in the same direction per ring, we have fiber-based recovery, because fibers are backed up by fibers. Extending this notion to WDM-based loopback, each ring is both primary for certain wavelengths and secondary for the remaining wavelengths. For simplicity, let us again consider just two wavelengths. Figure 9 shows that we cannot assign primary and secondary wavelengths in such a way that a wavelength is secondary or primary over a whole ring.

The use of double-cycle covers can also lead to asymmetric restoration times for a bidirectional connection. In particular, the links and nodes used to recover traffic crossing a link often depend on the direction of the traffic, with each direction being recovered by a separate cycle. The two directions on a link thus have different restoration times and timing jitter, which can lead to

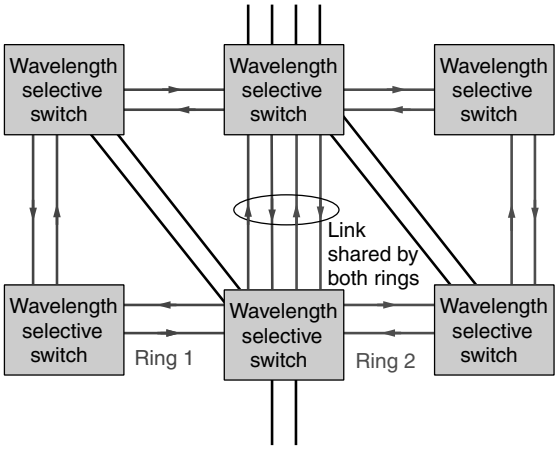


Figure 7. Two rings traversing separate resources over the same link.

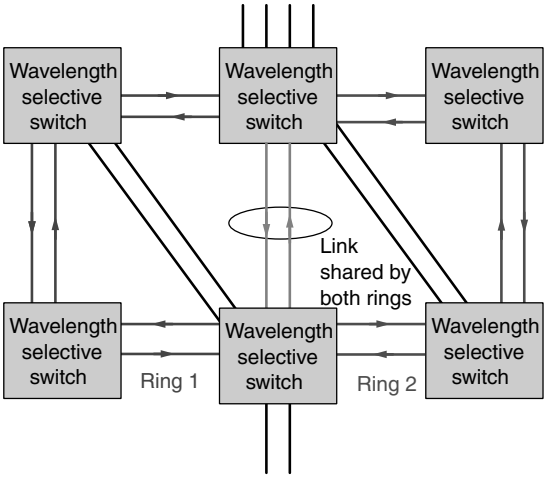


Figure 8. Two rings traversing shared resources over the same link.

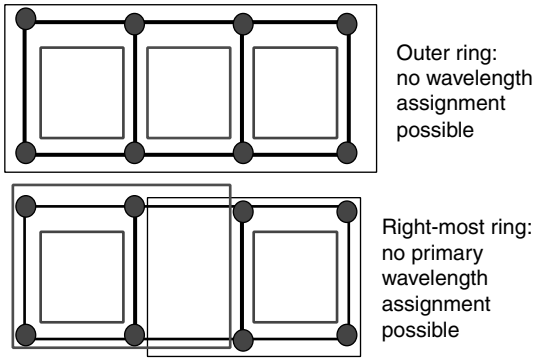


Figure 9. Example showing the problems of applying double cycle covers to wavelength recovery.

problems for bidirectional connections. In contrast, both SHRs and generalized loopback (discussed later in the section) avoid these problems by protecting bidirectional traffic with unique, bidirectional restoration paths.

Cycle covers work well for link failures, but have drawback for recovery from node failures, particularly when failures occur at nodes that are shared by one than one link. While node recovery can be effected with double-cycle ring covers, such restoration requires cumbersome hopping among rings. Moreover, if a link or node is added to a network, the cover of cycles can change significantly, limiting the scalability of double cycle covers. These drawbacks are a general property of ring embeddings, and are already found in SONET networks.

In order to avoid the limitations of ring covers, an approach using preconfigured protection cycles, or p-cycles, is given by Grover and Stamatelakis [92]. A *p-cycle* is a cycle on a redundant mesh network. Links on the p-cycle are recovered by using the p-cycle as a conventional BLSR. Links not on the p-cycle are recovered by selecting, along the p-cycle, a path connecting the nodes at either end of the failed link. Some difficulty arises from the fact that several p-cycles may be required to cover a network, making management among p-cycles necessary. A single p-cycle may be insufficient because a Hamiltonian circuit might not exist, even in a two-connected graph. Even finding p-cycles that cover a large number of nodes may be difficult. Some results [93–95] and conjectures [96,97] exist concerning the length of maximal cycles in two-connected graphs. The p-cycle approach is in effect a hybrid ring approach, which mixes link protection (for links not on the p-cycle) with ring recovery (for links on the p-cycle).

Another approach to link restoration on mesh networks, which we term generalized loopback, was first presented in 1999 [98]. The principle behind generalized loopback is to select a directed graph, called the *primary graph*, such that another directed graph, called the *secondary*, can be used to carry backup traffic for any link failure in the primary. Construction of a primary involves selection of a single direction for each link in the network. Loopback then occurs along the secondary graph in a manner akin to SONET BLSR. Figure 10 demonstrates generalized loopback for a simple network. In the figure, only two fibers per link are shown — one primary fiber and its corresponding secondary fiber. When the link $[Y, X]$ fails, traffic from the primary digraph floods onto the secondary digraph starting at Y. The secondary digraph carries this backup traffic from one endpoint of the failed node to the other endpoint, possibly along multiple paths. When traffic reaches X (along the first successful path), it is again placed on the primary fiber, as though no failure had occurred. Unnecessary backup paths are subsequently torn down. The fact that multiple paths may exist for restoration allows us to reclaim some arcs (fibers) from secondary digraphs to carry additional traffic. The capacity efficiency obtained in this manner is, for typical networks, in the order of 20% over methods, such as double-cycle cover, that require half of the network capacity to be devoted to recovery.

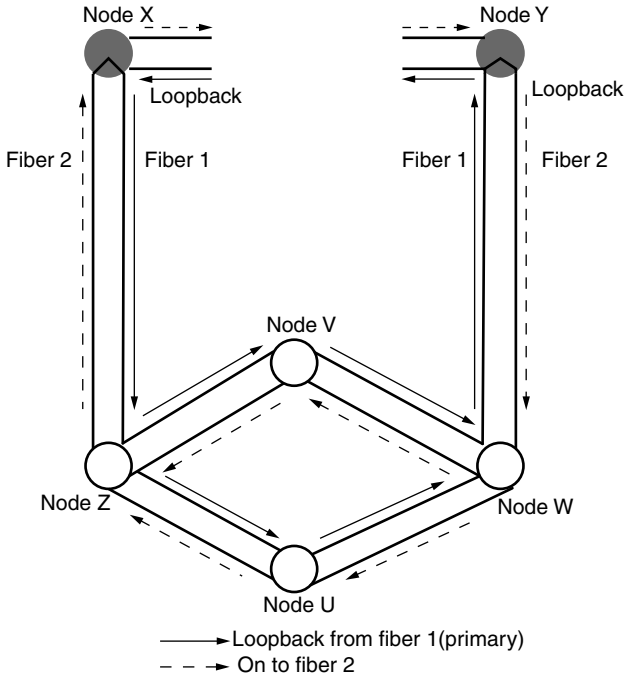


Figure 10. Generalized loopback.

7. PACKET-BASED APPROACHES

This section looks more closely at techniques particular to packet-based networks, giving more details for the failure detection techniques mentioned in Section 2 and the recovery schemes used when failures are detected. Some packet-based networks, such as FDDI, are based on ring topologies and use failure detection and recovery techniques nearly identical to those described in previous sections. For packet networks built with redundant mesh topologies, the approach is substantially different.

One protocol of particular interest and importance was developed in the mid-1980s [99] to allow redundant interconnection of LANs for reliability while avoiding problems of infinite routing loops. The model, known as an extended LAN, uses bridges to connect LANs and to forward traffic between LANs as necessary. The bridges cooperate in a distributed fashion to select a minimum spanning tree (MST) from the full-connectivity graph. All traffic is then routed along the MST, preventing cycles. Periodic configuration messages are sent from the root of the tree and forwarded over all LANs. Failure detection then relies on timeout mechanisms — if a bridge fails to hear the configuration message on any LAN to which it is attached, it acts to find a new route from the spanning tree to that LAN. The simplest method is to restart the search entirely, but some optimizations are possible.

The MST approach serves its purpose, allowing redundant topologies in packet-switched networks without allowing for routing loops. However, restricting traffic to flow along a tree can severely limit the capacity of the extended LAN, and can force traffic between two LANs that are close in the original mesh to follow a long path through the tree, consuming some of the capacity of many other LANs in the process.

Autonet, developed in the early 1990s, solved this problem to some extent by allowing the use of all links in the network [100]. Routing cycles and deadlocks in Autonet were prevented through the use of up*/down* (read "up star, down star," and alluding to regular expressions) routes. With this approach, all routers are assigned a unique number, and packets between two hosts in the network must follow a route that moves in a monotone sequence upward followed by a monotone sequence downward before exiting the network. Any route obeying this constraint can be used. Consider a cycle or self-cycle in routes, and label each link in the cycle as either up or down, depending on the identifiers assigned to the routers at the end of the link. Obviously, a cycle must contain both up and down links, and in particular must contain a two-link section with the first link down and the second up. No route can legally follow both links in this section, however, implying that deadlocks are impossible; thus, all up*/down* routes are mutually deadlock-free. Autonet relied on timeouts built into the switch hardware for detection of failed links and nodes, but was otherwise quite similar to the extended LAN approach in terms of reliability.

Since the early 1990s, many vendors of packet-based networks have recognized the importance of redundancy, and have introduced hardware and software support for combining physical channels into single logical channels between high-end switches. While this approach may seem fairly natural in a packet-based network, in which utilization is already based on statistical multiplexing of the links, some complexities must be addressed. These complexities arise from an assumption by higher-level protocols, in particular the Transmission Control Protocol (TCP), which packets sent through a network arrive in order of transmission. Use of multiple physical routes to carry packets from a single TCP connection often violates this assumption, causing TCP's congestion control mechanisms to drastically cut bandwidth. To address this issue, link aggregation schemes try to restrict individual TCP connections to specific links, and rely on the availability of many connections to provide good load balancing and capacity benefits from aggregation. Failure detection, as with Autonet, is generally handled in hardware, and results in routing reorganization. Unlike many backbone networks, the capacity of most packet-switched networks degrades in the presence of failures, encouraging network architects and managers to operate in somewhat risky modes in which inadequate capacity remains available after certain failures. This phenomenon can be observed even in high-end packet-based systems, including some SAN's backing bank operations.

In the wide area, fault tolerance in packet-switched networks relies on a combination of physical layer notification and rerouting by higher-level routing protocols. The Border Gateway Protocol (BGP) [101], a peer protocol to the Internet Protocol (IP), defines the rules for advertising and selecting routes to networks in the Internet. More specifically, it defines a homogeneous set of rules for interactions between *autonomous systems* (ASs), networks controlled by a single administrative entity. With each AS, administrators are free to select whatever routing protocol suits their fancy, but AS's must interact with each other in a standard way, as defined by BGP. BGP explicitly propagates failure information in the form of withdrawn routes, which cancel previously advertised routes to specific networks.

From the point of view of recovery in the context of optical backbone networks, the overhead required for packet-based systems depends critically on what functionalities are implemented in the optical domain. Restoration, in which, after a failure, excess bandwidth is claimed for the purpose of providing alternate routes to traffic around a failure, is challenging in the optical domain, since it requires operating on the whole datastream and possibly separating packets from a stream. In order to avoid packet-level operations, flow switching on a stream-by-stream basis, for instance using multiprotocol label switching (MPLS), is a promising alternative. Such stream-based operations are more amenable to optical processing. For recovery, stream-based processing reduces roughly to circuit-based recovery.

Packet-switched approaches for optical access seek to perform some subset of the functionalities of traditional opto-electronic packet-based networks optically [102]. These functionalities may be header recognition, buffering, packet insertion, packet reading, packet retrieval, and rate conversion. Performing such operations in the optical domain is challenging and no consensus has emerged regarding implementation. However, certain general statements can be made. Operations, such as buffering a stream, that involve significant timing issues or that introduce loss and distortion on the datastream, tend to be challenging. Replicating a stream, for instance, can be done using passive optical splitters and is therefore relatively straightforward. Merging streams, on the other hand, is challenging because of timing issues. The most challenging operations are the ones performed at the packet level. Again, different levels of difficulty arise. Reading signals from an optical datastream is possible by removing a fraction of the signal power and operating on that fraction. Retrieving a packet (reading the packet and removing it from the stream) is difficult because it involves performing an operation on the whole stream, as well as timing, phase, and polarization issues. Thus, operations such as packet-switching are also challenging because of issues of timing and speed of optical switches. Thus, fully optical packet-switched systems replicating the entire operations of electronic systems are still distant.

8. HIGH-SPEED LANS

The vast majority of the proposed architectures for LANs consist of star topologies or of networks built from combinations of star topologies, in which some type of switch, router, or other type of hub, is placed in the center of a topology and each node is directly connected to the hub [103]. The emergent 10 Gb/s standard (IEEE 802.3ae) for LANs and MANs also allows for optical stars and trees. From the point of reliability, stars present many weaknesses. In particular, a failure at the hub may entail failure of the whole network. However, other failures may occur even without outright failure of the hub. If the hub passively broadcasts, total failure of the hub is unlikely.

However, many partial failure scenarios exist: amplifier failures; port connection failures, at the access nodes or at the hub; transmitter or receiver failures at access nodes, for instance, because of laser failures; or cabling failures in the fiber itself. Such failures entail the failure of one or more arms of the star.

The center of a star topology is inherently a single point of failure, making complete replication of the system necessary to support recovery. Operation of a fully redundant system is difficult, however, as illustrated by existing reliable networks based on star topologies. Many enterprise networks and storage area networks (SANs), for example, are built as stars.

Such systems typically use single-wavelength optical connections rather than WDM, and rely on electronic switching. Enterprise networks are usually based on Gigabit Ethernet (GigE), while SAN's are based on Fibre Channel (FC). Network interface cards (NIC's), housing both a receiver and a transmitter, are optically connected to an electronic switch. The switch is closer to a traditional router than to the passive broadcast hubs or wavelength-selective switches discussed in the context of star-based WDM LANs. For such networks, redundancy is obtained by full duplication of all resources, as shown in Fig. 11.

In addition to replication, the two switches must be connected. Consider the case of failure of the primary NIC in server 1. Server 1 communicates via the secondary switch. Requiring other servers also to communicate via the secondary switch is undesirable. Indeed, although we show just two servers, such networks typically have many servers connected to them and reconfiguring so many connections simultaneously is difficult. Moreover, there is some delay involved in creating new connections through switch 2 owing to initialization overheads. To avoid reconfiguration at all servers, all servers other than server 1 continue to communicate with the primary switch and the two switches communicate with each other via the interswitch connection.

In the context of optical networks, an interswitch connection translates into connection between two hubs. In order to manage such an interhub connection, the hub needs to be equipped with far greater capabilities than simple optical broadcasting. Thus, it would appear that optical star dedicated networks will be difficult to deploy and that the means of providing robustness available in traditional star topologies cannot be easily extended to optical access networks.

The star topology for LAN's connected to a backbone is not limited to optical applications. Such an architecture has been proposed, for instance, for the Integrated Services LAN (ISLAN) defined by IEEE 802.9 using unshielded twisted pair.

While stars and topologies built from stars dominate in the LANs, LANs are also built using bus schemes. Bus schemes allow nodes to place and retrieve traffic using a shared medium. Figure 12 shows a folded bus and a dual bus. In a folded bus, a single bus, originating at a head end, serves all nodes. Typically, nodes use the bus first as a collection bus, onto which they place traffic (in the left-to-right direction in Fig. 12a). The last node folds back the bus to make it travel in the right-to-left direction. In the right-to-left direction, nodes collect traffic placed onto the bus. The traffic may be read only or read and removed. In the dual-bus architecture, two buses are used, each with its own headend. Folded and dual buses are simple options for LANs and certain types of MANs. In particular, they offer an effective way of sharing bandwidth among several users and are therefore attractive to allow nodes to access optical bandwidth, whether for a full fiber, a few wavelengths, or a single wavelength.

Folded and dual buses suffer from reliability drawbacks. Figure 13 shows a folded bus and a dual bus after a failure. Partial recovery can be effected by creating a bus on either side of the failure. For a dual-bus architecture, the node immediately upstream of the failure needs to be

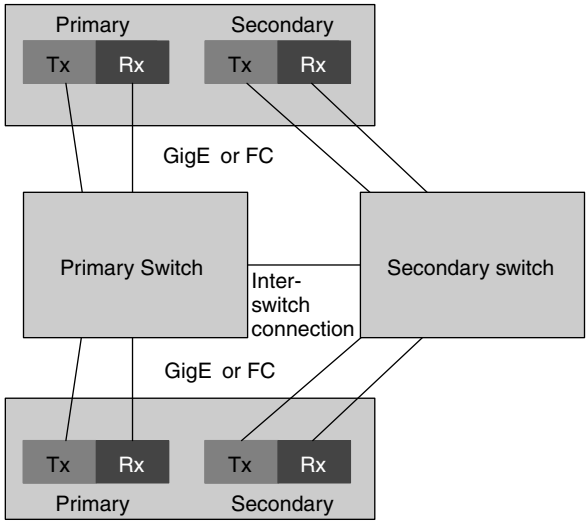


Figure 11. Redundant architecture in an enterprise network or LAN using traditional star topologies.

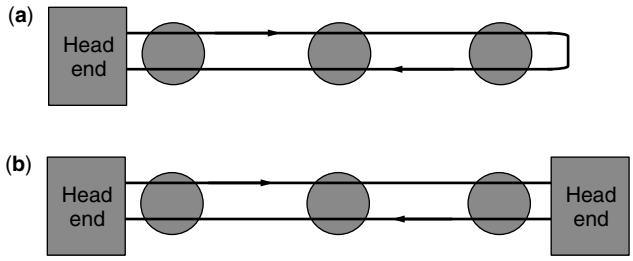


Figure 12. Folded (**a**) and dual (**b**) buses.

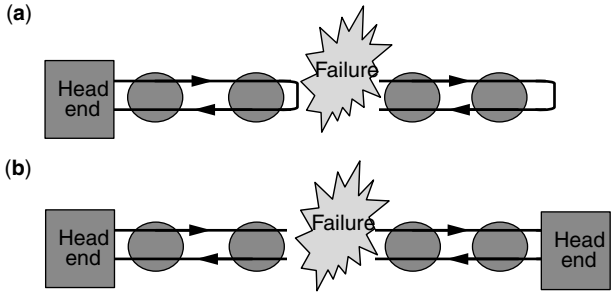


Figure 13. Folded (**a**) and dual (**b**) buses being restored as two folded buses and two dual buses, respectively, after a failure.

able to fold the bus. In order to reestablish full connectivity after a failure, the end nodes of the original buses must be able to connect outside the original buses to transmit traffic that was destined to traverse the cut.

BIOGRAPHY

Muriel Medard is an Assistant Professor in the Electrical Engineering and Computer Science (EECS) at MIT and a member of the Laboratory for Information and Decision Systems. She was previously an Assistant Professor at the Electrical and Computer Engineering Department and a member of the Coordinated Science Laboratory at the University of Illinois Urbana — Champaign. From 1995 to 1998 she was a Staff Member at MIT Lincoln Laboratory in the Optical Communications and the Advanced Networking Groups. Professor Medard received B.S. degrees in EECS and Mathematics in 1989, a B.S. degree in Humanities in 1990, a M.S. degree in Electrical Engineering 1991, and a Sc.D. degree in Electrical Engineering in 1995, all from the Massachusetts Institute of Technology (MIT), Cambridge. Medard's research interests are in the areas of reliable communications, particularly for optical and wireless networks. She received a 2001 NSF Career award. She was awarded the IEEE Leon K. Kirchmayer Prize Paper Award 2002 for her paper, "The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel."

BIBLIOGRAPHY

1. T. Anderson et al., *Dependability: Basic Concepts and Terminology*, Springer-Verlag, Wien (Vienna), Austria, 1992.
2. C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, Delayed Internet routing convergence, *Proc. ACM SIGCOMM Conf.*, 2000, pp. 175–187.
3. C. Labovitz, A. Ahuja, R. Watterhofer, and S. Venkatachary, The impact of Internet policy and topology on delayed routing convergence, *Proc. INFOCOM*, 2001.
4. K. Murakami and H. S. Kim, Virtual path routing for survivable ATM networks, *IEEE/ACM Trans. Network.* **4**: (1996).
5. D. J. Pai and H. L. Owen, An algorithm for bandwidth management with survivability constraints in ATM networks, *Proc. ICC*, 1997, Vol. 1, pp. 261–266.
6. O. Gerstel and R. Ramaswami, Optical layer survivability — an implementation perspective, *IEEE J. Select. Areas Commun.* 1885–1899 (2000).
7. D. Zhou and S. Subramanian, Survivability in optical networks, *IEEE Network* 16–23 (Nov./Dec. 2000).
8. C. J. Green, Protocols for a self-healing network, *Proc. Military Communications Conf.* (*MILCOM*), 1995, Vol. 1, pp. 252–256.
9. T. E. Stern and K. Bala, *Multiwavelength Optical Networks: A Layered Approach*, Prentice-Hall, Upper Saddle River, NJ, 2000.
10. A. Banerjea, C. J. Parris, and D. Ferrari, Recovering guaranteed performance service connections from single and multiple faults, *Proc. GLOBECOM*, 1994, Vol. 1, pp. 162–166.
11. R. Doverspike, A multi-layered model for survivability in intra-LATA transport networks, *Proc. GLOBECOM*, 1991, pp. 2025–2031.
12. T.-H. Wu, *Fiber Network Service Survivability*, Artech House, 1992.
13. D. Johnson et al., Distributed restoration strategies in telecommunications networks, *Proc. IEEE Int. Conf. Communications*, 1994, Vol. 1, pp. 483–488.
14. R. Nakamura, H. Ono, and K. Nishikawara, Reliable switching services, *Proc. GLOBECOM*, 1994, Vol. 3, pp. 1596–1600.
15. J. Veerasamy, S. Vemkatesan, and J. C. Shah, Effect of traffic splitting on link and path restoration planning, *Proc. GLOBECOM*, 1994, Vol. 3, pp. 1867–1871.
16. M. Tomizawa, Y. Yamabayashi, N. Kawase, and Y. Kobayashi, Self-healing algorithm for logical mesh connection on ring networks, **30**: 1615–1616 (Sept. 15 1994).
17. J. Sosnosky, Service application for sonet dcs distributed restoration, **12**: 59–68 (1994).
18. D. Edinger, P. Duthie, and G. R. Prabhakara. A new answer to fiber protection. 53–55, April 9, 1990.
19. T.-H. Wu and W. I. Way, A novel passive protected sonet bidirectional self-healing ring architecture, *jtl* **10**: (Sept. 1992).
20. W. D. Grover, The selfhealing^{TM} network, *Proc. GLOBECOM*, 1987, pp. 1090–1095.
21. C. H. Yang and S. Hasegawa, Fitness: Failure immunization technology for network service survivability, *Proc. GLOBECOM*, 1988, Vol. 3, pp. 47.3.1–47.3.6.
22. H. Fujii and N. Yoshikai, Double search self-healing algorithm and its characteristics, **77**: 975–995 (1994).
23. H. Sakauchi, Y. Okanoue, H. Okazaki, and S. Hasegawa, Distributed self-healing control in SONET, *J. Network Syst. Manage.* **1**(2): 123–141 (1993).
24. T.-H. Wu, A passive protected self-healing mesh network architecture and applications, *IEEE/ACM Trans. Network.* 40–52 (Feb. 1994).
25. H. Kobrinski and M. Azuma, Distributed control algorithms for dynamic restoration in dcs mesh networks: Performance evaluation, *Proc. GLOBECOM*, 1993, Vol. 3, pp. 1584–1588.
26. A. Gersht, S. Kheradpir, and A. Shulman, Dynamic bandwidth-allocation and path-restoration in SONET self-healing networks, *IEEE Trans. Reliability* **45**: 321–331 (June 1996).
27. M. Barezzani, E. Pedrinelli, and M. Gerla, Protection planning in transmission networks, *Proc. ICC*, 1992, pp. 316.4.1–316.4.5.
28. C. E. Chow, J. Bicknell, S. McCaughey, and S. Syed, A fast distributed network restoration algorithm, *Proc. 12th Int. Phoenix Conf. Computers and Communications*, March 1993, Vol. 1, pp. 261–267.
29. J. Bicknell, C. E. Chow, and S. Syed, Performance analysis of fast distributed network restoration algorithms, *Proc. IEEE GLOBECOM*, 1993, Vol. 3, pp. 1596–1600.
30. R. Doverspike and B. Wilson, Comparison of capacity efficiency of dcs network restoration routing techniques, volume 2, 1994.
31. T. Frisanco, Optimal spare capacity design for various protection switching methods in atm networks, *Proc. ICC*, 1997, Vol. 1, pp. 293–298.

32. N. Wauters, B. Van Caenegem, and P. Demeester, Spare capacity assignment for different restoration strategies in mesh survivable networks, *Proc. ICC*, 1997, Vol. 1, pp. 288–292.
33. K. Menger, *Zur allgemeinen Kurventheorie*, Fundamenta Mathematicae, 1927.
34. M. Stoer, *Design of Survivable Networks*, Springer-Verlag, 1992.
35. R. Bhandari, Optimal diverse routing in telecommunication fiber networks, *Proc. IEEE INFOCOM*, May 1994, Vol. 3, pp. 11.c.3.1–11.c.3.11.
36. S. Z. Shaikh, Span-disjoint paths for physical diversity in networks, *Proc. IEEE Symp. Computers and Communications*, 1995, pp. 127–133.
37. J. W. Suurballe, Disjoint paths in a network, pages 125–145, 1974.
38. W. T. Zaumen and J. J. Garcia-Luna Aceves, Dynamics of distributed shortest-path routing algorithms, *Proc. 21st SIGCOMM Conf.*, Sept. 3–6, 1991, ACM Press, 1991, Vol. 21, pp. 31–43.
39. J. S. Whalen and J. Kenney, Finding maximal link disjoint paths in a multigraph, *Proc. GLOBECOM*, 1990, pp. 403.6.1–403.6.5.
40. P. Mateti and N. Deo, On algorithms for enumerating all circuits of a graph, **5**: (March 1976).
41. A. Itai and M. Rodeh, The multi-tree approach to reliability in distributed networks, Number 79, 1988.
42. M. Médard, S. G. Finn, R. G. Gallager, and R. A. Barry, Redundant trees for automatic protection switching in arbitrary node-redundant or edge-redundant graphs, *Proc. ICC*, 1998.
43. R. Kawamura, H. Hadama, and I. Tokizawa, Implementation of self-healing function in ATM networks, *J. Network Syst. Manage.* **3**(3): 243–264 (1995).
44. N. D. Lin, A. Zolfaghari, and B. Lusignan, ATM virtual path self-healing based on a new path restoration protocol, *Proc. GLOBECOM*, 1994, Vol. 2, pp. 794–798.
45. D. K. Hsing, B.-C. Cheng, G. Goncu, and L. Kant, A restoration methodology based on preplanned source routing in ATM networks, *Proc. ICC*, 1997, Vol. 1, pp. 277–282.
46. R. S. K. Chng et al., A multi-layer restoration strategy for reconfigurable networks, *Proc. GLOBECOM*, 1994, Vol. 3, pp. 1872–1878.
47. S. Hasegawa, A. Kanemasa, H. Sakaguchi, and R. Maruta, Dynamic reconfiguration of digital cross-connect systems with network control and management, *Proc. GLOBECOM*, 1994, pp. 28.3.2–28.3.5.
48. A. Gersht and S. Kheradpir, Real-time bandwidth allocation and path restorations in SONET-based self-healing mesh networks, *Proc. IEEE Int. Conf. Communications*, 1993, Vol. 1, pp. 250–255.
49. J. E. Baker, A distributed link restoration algorithm with robust preplanning, *Proc. IEEE GLOBECOM*, 1991, Vol. 1, pp. 10.4.1–10.4.6.
50. S. Ramamurthy and B. Mukherjee, Survivable WDM mesh networks, part I — protection, *Proc. IEEE INFOCOM*, 1999, pp. 744–751.
51. J. Anderson, B. T. Doshi, S. Dravida, and P. Harshavardhana, Fast restoration of ATM networks, *IEEE J. Select. Areas Commun.* **12**(1): 128–136 (Jan. 1994).
52. Y. Xiong and L. G. Mason, Restoration strategies and spare capacity requirements in self-healing ATM networks, *IEEE J. Lightwave Commun.* **7**(1): 98–110 (Feb. 1999).
53. M. Azuma et al., Network restoration algorithm for multimedia communication services and its performance characteristics, *IEICE Trans. Commun.* **E78-B**(7): 987–994 (July 1995).
54. R. Kawamura, K. Sato, and I. Tokizawa, High-speed self-healing techniques utilizing virtual paths, *Proc. 5th Int. Network Planning Symp.*, May 1992.
55. M. Boyden T.-H. Wu, and R. H. Caldwell, A multi-period design model for survivable network architecture selection for sdh/sonet interoffice networks, volume **40**: 417–432 (Oct. 1991).
56. O. J. Wasem, An algorithm for designing rings for survivable fiber networks, volume **40**: (1991).
57. T.-H. Wu and M. E. Burrows, Feasibility study of a high-speed sonet self-healing ring architecture in future interoffice networks, pages 33–51, Nov. 1990.
58. C.-C. Shyur, Y.-M. Wu, and C.-H. Chen, A capacity comparison for sonet self-healing ring networks, *Proc. GLOBECOM*, 1993, pp. 1574–1578.
59. J. B. Slevinsky, W. D. Grover, and M. H. MacGregor, An algorithm for survivable network design employing multiple self-healing rings, *Proc. GLOBECOM*, 1993, Vol. 3, pp. 1568–1573.
60. J. Shi and J. Fonseka, Interconnection of self-healing rings, *Proc. ICC*, 1996, Vol. 1.
61. C.-C. Shyur, S.-H. Tsao, and Y.-M. Wu. Survivable network planning methods and tools in Taiwan, Sept. 1995.
62. L. M. Gardner et al., Techniques for finding ring covers in survivable networks, *Proc. GLOBECOM*, 1994, Vol. 3.
63. T.-H. Wu, D. J. Kolar, and R. H. Cardwell, High-speed self-healing ring architectures for future interoffice networks, *Proc. GLOBECOM*, 1989, Vol. 2, pp. 23.1.–23.1.7.
64. E. L. Hahne and T. D. Todd, Fault-tolerant multimesh networks, *Proc. GLOBECOM*, 1992, pp. 627–632.
65. R. B. Magill, A bandwidth efficient self-healing ring for b-isdn, *Proc. ICC*, 1997.
66. Y. Kajiyama, N. Tokura, and K. Kikuchi, An atm vp-based self-healing ring, **12**: (Jan. 1994).
67. F. E. Ross, An overview of FDDI: The fiber distributed data interface, *IEEE J. Select. Areas Commun.* **7**: 1043–1051 (Sept. 1989).
68. F. E. Ross, Fiber distributed data interface: An overview, *Proc. 15th Conf. Local Computer Networks*, 1990, pp. 6–11.
69. R. O. LaMaire, FDDI performance at 1 Gbit/s, *Proc. IEEE Int. Conf. Communications*, 1991, pp. 174–183.
70. T. S. Jones and A. Louri, Media access protocols for a scalable optical interconnection network, May 1998.
71. M. A. Marsan et al., An almost optimal MACprotocol for all-optical WDM multi-rings with tunable transmitters and fixed receivers, *Proc. IEEE Int. Conf. Communications*, May 1997, Vol. 1, pp. 437–442.
72. M. A. Marsan et al., All-optical WDMmulti-rings with differentiated qos., *IEEE Commun. Mag.* **37**: 58–66 (Feb. 1999).
73. M. A. Marsan et al., SR/sup 3/:a bandwidth-reservation MACprotocol for multimedia applications over all-optical WDM multi-rings, *Proc. INFOCOM '97*, 1997, Vol. 2.

74. A. Bianco et al., A-posteriori access strategies in all-optical slotted WDM rings, *Proc. Global Telecommunications Conf.*

75. A. Louri and R. Gupta, Hierarchical optical interconnection network HORN: Scalable interconnection network for multiprocessors and multicomputers, Jan. 1997.

76. Resilient Packet Ring Working Group.

77. A. Banerjea, C. J. Parris, and D. Ferrari, Recovering guaranteed performance service connections from single and multiple faults, *Proc. GLOBECOM*, 1994, Vol. 1, pp. 162–166.

78. T. H. Wu, D. J. Kolar, and R. H. Cardwell, Survivable network architectures for broad-band fiber optic networks: Model and performance comparison, *IEEE J. Lightwave Commun.* **6**(11): (Nov. 1988).

79. K. T. Newport and P. K. Varshney, Design of survivable communication networks under performance constraints, *IEEE Trans. Reliability* **40**: 433–440 (Oct. 1991).

80. T.-H. Wu and S. Fouad Habiby, Strategies and technologies for planning a cost-effective survivable network architecture using optical switches, *IEEE Trans. Reliability* **8**(2): 152–159 (Feb. 1991).

81. R.-H. Jan, F.-J. Hwang, and S. T. Cheng, Topologocial optimization of a communication network subject to a reliability constraint, *IEEE Trans. Reliability* **42**(1): (March 1993).

82. W. D. Grover, Case studies of survivable ring, mesh and mesh-arc hybrid networks, *Proc. GLOBECOM*, 1992, pp. 633–638.

83. C. Thomassen, On the complexity of finding a minimum cycle cover of a graph, *SIAM J. Comput.* **26**: 675–677 (June 1997).

84. A. Itai, R. J. Lipton, C. H. Papadimitriou, and M. Rodeh, Covering graphs with simple circuits, *SIAM J. Comput.* **10**: 746–750 (1981).

85. G. Fan, Covering graphs by cycles, *SIAM J. Comput.* **5**: 491–496 (Nov. 1992).

86. G. Ellinas, T. E. Stern, and A. Hailemarian, Link failure restoration in optical networks with arbitrary mesh topologies and bi-directional links, 1997.

87. G. Ellinas and T. E. Stern, Automatic protection switching for link failures in optical networks with bi-directional links, *Proc. GLOBECOM*, 1996.

88. F. Jaeger, A survey of the double cycle cover conjecture, in *Cycles in Graphs*, Annals of Discrete Mathematics, Vol. 115, North-Holland, 1985.

89. P. D. Seymour, Sums of circuits, in U. S. R. Murty and J. A. Bondy, eds., *Graph Theory and Related Topics*, Academic Press, New York, 1979, pp. 341–355.

90. G. Szekeres, Polyhedral decomposition of cubic graphs, *J. Austral. Math. Soc.* **8**: 367–387 (1973).

91. L. Goddyn, A girth requirement for the double cycle cover conjecture, in *Cycles in Graphs*, Annals of Discrete Mathematics, Vol. 115, North-Holland, 1985, pp. 13–26.

92. W. D. Grover and D. Stamatelakis, Cycle-oriented distributed preconfiguration: Ring-like speed with mesh-like capacity for self-planning network reconfiguration, *Proc. IEEE Int. Conf. Communications*, 1998, Vol. 2, pp. 537–543.

93. I. Fournier, Longest cycles in 2-connected graphs of independence number α, in *Cycles in Graphs*, Annals of Discrete Mathematics, Vol. 115, North-Holland, 1985, pp. 201–204.

94. B. Jackson, Hamilton cycles in regular 2-connected graphs, *J. Comb. Theory Ser. B* **29**: 27–46 (1980).

95. Y. Zhu, Z. Liu, and Z. Yu, An improvement of Jackson's result on Hamilton cycles in 2-connected graphs, in *Cycles in Graphs*, Annals of Discrete Mathematics, Vol. 115, North-Holland, 1985, pp. 237–247.

96. R. H$\acute{}$aggkvist and B. Jackson, A note on maximal cycles in 2-connected graphs, in *Cycles in Graphs*, Annals of Discrete Mathematics, Vol. 115, North-Holland, 1985, pp. 205–208.

97. D. R. Woodall, Maximal circuits of graphs II, *Studia Sci. Math. Hungar.* **10**: 103–109 (1975).

98. M. Médard, S. G. Finn, and R. A. Barry, Wdm loopback recovery in mesh networks, *Proc. IEEE INFOCOM*, 1999.

99. R. Perlman, An algorithm for distributed computation of a spanning tree in an extended LAN, *Proc. 9th Symp. Data Communications,* (*SIGCOMM'85*), Whistler Mountain, British Columbia, Canada, 1985, pp. 44–53.

100. M. D. Schroeder et al., Autonet: A high-speed, self-configuring local area network using point-to-point links, *IEEE J. Select. Areas Commun.* **9**: 1318–1335 (Oct. 1991).

101. Y. Rekhter and T. Li, *A Border Gateway Protocol 4* (*BGP-4*), Internet Engineering Task Force RFC 1771, March 1995.

102. E. Modiano, Wdm-based packet networks, *IEEE Commun. Mag.* **37**: 130–135 (March 1999).

103. P. A. Humblet, R. Ramaswami, and K. N. Sivarajan, An efficient communication protocol for high-speed packet-switched multichannel networks, *IEEE J. Select. Areas Commun.* **11**: 568–578 (May 1993).

NETWORK SECURITY

ROLF OPPLIGER
eSECURITY Technologies
Rolf Oppliger
Bern, Switzerland

1. INTRODUCTION

According to Shirey [1], the term *computer network* (or *network*) refers to, "a collection of host computers together with the subnetwork or internetwork through which they can exchange data." Many different technologies and communication protocols can be used (and are in use) to build and operate computer networks. Examples include the IEEE 802 family of protocols for local area networking, the Point-to-Point Protocol (PPP) for dialup networking, and the TCP/IP protocol suite for internetworking.[1]

Almost all contemporary networking technologies and communication protocols are highly complex and have not been designed with security in mind. Consequently, they

[1] The acronym TCP/IP refers to an entire suite of communications protocols that center around the Transmission Control Protocol (TCP) and the Internet Protocol (IP). The emerging use of TCP/IP networking has led to a global system of interconnected hosts and networks that is commonly referred to as the Internet.

are inherently vulnerable and exposed to a variety of threats and corresponding attacks. To make things worse, there are a number of reasons why networked computer systems are inherently more vulnerable and exposed to threats and corresponding attacks than their standalone counterparts:

- More points exist from where an attack can be launched. Note that a computer system that is inaccessible or unconnectable to users cannot be attacked. Consequently, by adding more network connections (i.e., network connectivity) for legitimate users, more possibilities to attack the system are automatically added, as well.
- The physical perimeter of a networked computer system is artificially extended by having it connect to a computer network. This extension typically leads beyond what is actually controllable by a system administrator.
- Networked computer systems typically run software that is inherently more complex and error-prone. There are many network software packages that have a long bug record and that are known to be "buggy" accordingly (e.g., the UNIX sendmail daemon). More often than not, intruders learn about these bugs before system administrators do. To make things worse, intruders must know and be able to exploit one single bug, whereas system administrators must know and be able to fix all of them. Consequently, the workload between system administrators and potential intruders is asymmetrically distributed.

In essence, the aim of network security is to provide the technologies, mechanisms, and services that are required and can be used to protect the computational and networking resources against accidental and/or intentional threats. Mainly because of the importance of computer networks in daily life, network security is a hot topic today.

This article provides an overview and discussion about the current state -of- the -art and future perspectives in network security in general, and Internet security in particular. As such, it is organized as follows. Possible threats and attacks against computer networks and distributed systems are overviewed and briefly discussed in Section 2. The OSI security architecture is introduced in Section 3. The architecture provides a useful terminology that is applicable to a wide range of networking technologies and corresponding communication protocols. As a case study, Internet security is further addressed in Section 4. Finally, conclusions are drawn and an outlook is given in Section 5. Some parts of this article are taken from Ref. 2. Readers may refer to this reference book to get some further information about network security in general, and Internet security in particular.

2. THREATS AND ATTACKS

A threat refers to "a potential for violation of security, which exists when there is a circumstance, capability, action, or event that could breach security and cause harm [1]. That is, a threat is a possible danger that might exploit a vulnerability." A threat can be either accidental (e.g., caused by a natural disaster) or intentional (e.g., an attack). In the sequel, we focus only on intentional threats and corresponding attacks that may be launched either by legitimate users (i.e., insiders) or — more importantly — outside attackers (i.e., outsiders). All statistical investigations reveal (or confirm) the fact that most attacks are launched by insiders rather than outsiders. This is because insiders generally have more knowledge and possibilities to attack computer systems that store, process or transmit valuable information assets.

Again referring to Shirey [1], an attack is "an assault on system security that derives from an intelligent threat, i.e., an intelligent act that is a deliberate attempt (especially in the sense of a method or technique) to evade security services and violate the security policy of a system." There are many attacks that can be launched against computer networks and the systems they interconnect. Most attacks are due to vulnerabilities in the underlying network operating systems. In fact, the complexity of contemporary network operating systems makes it possible and very likely that we will see an increasingly large number of network-based and software-driven attacks in the future. What we experience today with macro viruses and network worms is only the tip of an iceberg.

With regard to telecommunications, it is common to distinguish between passive and active attacks:

- A *passive attack* attempts to learn or make use of information but does not affect system or network resources.
- An *active attack* attempts to alter system or network resources and affect their operation.

Passive and active attacks are typically combined to more effectively invade a computing or networking environment. For example, a passive wiretapping attack can be used to eavesdrop on authentication information that is transmitted in the clear (e.g., a username and password), whereas this information can later be used to masquerade another user and to actively attack the corresponding computer system. Passive and active attacks are further explored next.

2.1. Passive Attacks

As mentioned above, a passive attacker attempts to learn or make use of information but does not affect system or network resources. As such, a passive attack primarily threatens the confidentiality of data being transmitted. This data may include anything, including, for example, confidential electronic mail messages or usernames and passwords transmitted in the clear. In fact, the cleartext transmission of authentication information is the single most important vulnerability in computer networks today.

In regard to the intruder's opportunities to interpret and extract the information that is encoded in the transmitted data, it is common to distinguish between passive wiretapping and traffic analysis attacks:

- In a *passive wiretapping* attack, the intruder is able to interpret and extract the information that is encoded in the transmitted data. For example, if two parties communicate unencrypted, a passive wiretapper is trivially able to extract all information that is encoded in the data.
- In a *traffic analysis* attack, the intruder is not able to interpret and extract the information that the transmitted data encodes (because, e.g., the information is encrypted for transmission). Instead, *traffic analysis* refers to the inference of information from the observation of external traffic characteristics. For example, if an attacker observes that two companies — one financially strong, the other financially weak — begin to trade a large number of encrypted messages, he/she may infer that they are discussing a merger. Many other examples occur in military environments.

The feasibility of a passive attack primarily depends on the physical transmission media in use and their physical accessibility to potential intruders. For example, mobile communications is inherently easy to tap, whereas metallic transmission media at least require some sort of physical access. Lightwave conductors also can be tapped, but this is technically more challenging and expensive. Also note that the use of concentrating and multiplexing techniques, in general, makes it more difficult to passively attack data in transmission. Because of these difficulties, it is more likely that computer networks are passively attacked at the edge (e.g., local-area network segments that are connected to a wide-area network) than in its core or backbone.

It is, however, also important to note that a passive attacker does not necessarily have to tap a physical communications line. Most network interfaces can operate in a so-called promiscuous mode. In this mode, they are able to capture all frames transmitted on the local area network segment they are connected to, rather than just the frames addressed to the computer systems of which they are part. This capability has many useful purposes for network analysis, testing, and debugging (e.g., by utilities such as etherfind and tcpdump in the case of the UNIX operating system). Unfortunately, the capability also can be used by attackers to snoop on all traffic on a particular network segment. Several software packages are available for monitoring network traffic, primarily for the purpose of network management. These software packages are dual-use, meaning they can, for example, be effective in eavesdropping and capturing email messages or usernames and passwords as they are transmitted over shared media and communication lines.

A number of technologies can be used to protect a network environment against passive wiretapping attacks. For example, switched networks are more difficult to wiretap (because data are not broadcast and sent to all potential recipients). Consequently, the use of switched networks in the local area has had a very positive effect on the possibility to passively attack computer networks. Also, a few tools attempt to detect network interfaces that operate in promiscuous mode. For example, a tool named AntiSniff[2] implements a number of tricks to do so. One trick is to send an Ethernet frame with an invalid MAC address to a system and to encapsulate an Internet Control Message Protocol (ICMP) request packet with a valid IP header in the Ethernet frame (ICMP is the control protocol that complements IP). If the targeted system has a network interface that operates in promiscuous mode, it will grab the frame from the Ethernet segment decapsulate it and properly forward it to the local IP module. The IP module, in turn, will decapsulate and receive the ICMP echo request and eventually return a corresponding ICMP response. As a consequence, the targeted system reacts on something it should not have reacted (i.e., because of the invalid MAC address it should not have received the ICMP request in the first place). Obviously, it is simple to hide a network interface that operates in a promiscuous mode simply by not responding to ICMP requests that are encapsulated in Ethernet frames with invalid MAC addresses. What we are going to see in the future is that tools that can be used to passively wiretap network segments and tools that try to detect these tools play "hide and seek" on network segments. Last but not least, the use of data encryption is both effective and efficient against passive wiretapping attacks. In fact, it is the preferred technology and the technology of choice for network practitioners.

Contrary to passive wiretapping attacks, protection against traffic analyses is much more complicated and requires more sophisticated security technologies. Note that the use of encryption techniques does not protect against traffic analysis attacks. In fact, there are only a few technologies readily available to protect against traffic analysis attacks. Exemplary technologies include traffic padding (as discussed later) and a few privacy-enhancing technologies (PETs), such as onion routing (not addressed in this article). There is a lot of room for further research and development in this area.

2.2. Active Attacks

As mentioned above, an active attacker attempts to alter system or network resources and affect their operation. Consequently, an active attack primarily threatens the integrity or availability of data being transmitted. What this basically means is that the attacker can modify, extend, delete, or replay data units that are transmitted in computer networks and distributed systems.

The underlying reason why most active attacks are possible and fairly easy to launch in computer networks and distributed systems is that the data units that are sent and received are seldom protected in terms of authenticity and integrity. Examples of such data units include Ethernet frames, IP packets, User Datagram Protocol (UDP) datagrams, and Transmission Control Protocol (TCP) segments. Consequently, it is simple to do such things as flooding a recipient and cause a "denial of service" or "degradation of service," spoofing the source of data units or the identity of somebody else, or taking over and "hijacking" established network connections. Active attacks are very powerful and it is possible and very likely that we

[2] *http://www.securitysoftwaretech.com/antisniff/*.

will see many other active attacks being discovered and published in the future.

A number of technologies can be used to protect against some active attacks. Most of these technologies use cryptographic techniques to protect the authenticity and integrity of data units that are transmitted. There are, however, also a number of active attacks that are hard to protect against. Examples include denial-of-service and degradation-of-service attacks, as well as their distributed counterparts (i.e., distributed denial-of-service and degradation-of-service attacks). Similar to the real world, protection against this kind of attacks is very difficult to achieve in the digital world of computer networks. How would you, for example, protect your mailbox in the real world against somebody who fills it up with empty paper sheets? There seems to be no simple answer to this question, and the problem is difficult to address in either the real or digital world. In the digital world the problem is even more worrisome, simply because the corresponding attacks are much simpler (and less expensive) to launch. There are, for example, many tools that automatically fill up the mailboxes of particular victims.

3. OSI SECURITY ARCHITECTURE

According to Shirey [1], a *security architecture* refers to "a plan and set of principles that describe (a) the security services that a system is required to provide to meet the needs of its users, (b) the system elements required to implement the services, and (c) the performance levels required in the elements to deal with the threat environment." As such, a security architecture is the result of applying good principles of systems engineering and addresses issues related to physical security, computer security, communication security, organizational security (e.g., administrative and personnel security), and legal security. This is complicated and difficult to achieve, but it is very important. More often than not, systems and applications are designed, implemented and deployed without having an appropriate security architecture in mind.[3]

To extend the field of application of the reference model for open systems interconnection (OSI), the ISO/IEC JTC1 appended a security architecture as part two of ISO/IEC 7498 in the late 1980s [3]. The OSI security architecture is still valid and in use today. It provides a general description of security services and related security mechanisms, which may be provided by a computer network, and defines the positions within the OSI reference model where the services and mechanisms may be provided. Since its publication, the OSI security architecture has turned out to be a primary reference for network security professionals. In 1991, the ITU-T adopted the OSI security architecture in its recommendation X.800 [4] and the Privacy and Security Research Group (PSRG) of the Internet Research Task Force (IRTF) adopted the OSI security architecture in a corresponding Internet security architecture[4] in the early 1990s. In essence, ISO/IEC 7498-2, ITU-T X.800, and the Internet security architecture describe the same security architecture, and in this article we use the term OSI security architecture to collectively refer to all of them.

In short, the OSI security architecture provides a general description of security services and related security mechanisms and discusses their interrelationships. It also shows how the security services map onto a given network architecture and briefly discusses their appropriate placement within the OSI reference model. Having the abovementioned definition of a security architecture in mind, it is obvious that the OSI security architecture does not conform to it. In fact, the OSI security architecture rather refers to a (terminological) framework and a general description of security services and related security mechanisms than to a full-fledged security architecture. Nevertheless, we use it in this article to serve as a starting point for subsequent discussions.

3.1. Security Services

The OSI security architecture distinguishes among five complementary classes of security services. These classes comprise authentication, access control, data confidentiality, data integrity, and nonrepudiation services. Just as layers define functionality in the OSI reference model, so do security services in the OSI security architecture define various security objectives and aspects relevant for computer networks and distributed systems.

- *Authentication services* provide for the authentication of communicating peers or data origins:
  - A *peer entity authentication service* provides the ability to verify that a peer entity in an association is the one it claims to be. In particular, a peer entity authentication service provides assurance that an entity is not attempting to masquerade or perform an unauthorized replay of some previous association. Peer entity authentication is typically performed either during a connection establishment phase or, occasionally, during a data transfer phase.
  - A *data origin authentication service* allows the sources of data received to be verified to be as claimed. A data origin authentication service, however, cannot provide protection against the duplication or modification of data units. In this case, a data integrity service must be used in conjunction with a data origin authentication service. Data origin authentication is typically provided during a data transfer phase.

  Authentication services are important because they are a prerequisite for proper authorization, access control, and accountability. *Authorization* refers to the process of granting rights, which includes the granting of access based on access rights. Access control refers to the process of enforcing access rights, and accountability to the property that ensures that the actions of a principal may be traced uniquely to this particular principal.

[3] Refer to *http://www.esecurity.ch/security_architectures.pdf* for a white paper that describes the role and importance of having an appropriate security architecture.

[4] This work has been abandoned.

- *Access control services* provide for the protection of system or network resources against unauthorized use. As mentioned above, access control services are often closely tied to authentication services; For instance, a user or a process acting on a user's behalf must be properly authenticated before an access control service can effectively mediate access to system resources. In general, access control services are the most commonly used services in both computer and communication security.
- *Data confidentiality* refers to the property that information is not made available or disclosed to unauthorized individuals, entities, or processes. Thus, *data confidentiality services* provide for the protection of data from unauthorized disclosure:
  - A *connection confidentiality service* provides confidentiality of all data transmitted in a connection.
  - A *connectionless confidentiality service* provides confidentiality of single data units.
  - A *selective field confidentiality service* provides confidentiality of only certain fields within the data during a connection or in a single data unit.
  - A *traffic flow confidentiality service* provides protection of information that may otherwise be compromised or indirectly derived from a traffic analysis.

  The provision of a traffic flow confidentiality service requires fundamentally different security mechanisms than the other data confidentiality services.
- *Data integrity* refers to the property that information is not altered or destroyed in some unauthorized way. Thus, *data integrity services* provide protection of data from unauthorized modifications:
  - A *connection integrity service with recovery* provides integrity of data in a connection. The loss of integrity is recovered, if possible.
  - A *connection integrity service without recovery* provides integrity of data in a connection. In this case, however, the loss of integrity is not recovered.
  - A *selected field connection integrity service* provides integrity of specific fields within the data during a connection.
  - A *connectionless integrity service* provides integrity of single data units.
  - A *selected field connectionless integrity service* provides integrity of specific fields within single data units.

  Note that on a connection, the use of a peer entity authentication service at the start of the connection and a connection integrity service during the connection can jointly provide for the corroboration of the source of all data units transferred on the connection, the integrity of those data units, and may additionally provide for the detection of duplication of data units, for example, by using sequence numbers.
- *Nonrepudiation services* prevent one of the entities involved in a communication from later denying having participated in all or part of the communication. Consequently, they have to provide some sort of protection against the originator of a message or action denying that he/she has originated the message or the action, as well as against the recipient of a message denying having received the message. Consequently, there are two non-repudiation services to be distinguished:
  - A *nonrepudiation service with proof of origin* provides the recipient of a message with a proof of origin.
  - A *nonrepudiation service with proof of delivery* provides the sender of a message with a proof of delivery.

  Nonrepudiation services are becoming increasingly important in the context of electronic commerce (e-commerce) on the Internet [5]. For example, a non-repudiation service with proof of delivery may be important for secure messaging (in addition to any secure messaging scheme that employs digital envelopes and digital signatures). The corresponding service is sometimes also referred to as "certified mail." Certified mail is certainly a missing piece for the more professional use electronic mail.

The security services mentioned in the OSI security architecture can be complemented by anonymity or pseudonymity services. These services are not addressed in this article. Sometimes, the availability of anonymity or pseudonymity services directly contradicts the availability of other security services, such as authentication and access control services.

In either case, a security service can be implemented by one or several security mechanisms. The security mechanisms that are addressed in the OSI security architecture are briefly overviewed next.

3.2. Security Mechanisms

The OSI security architecture distinguishes between specific security mechanisms and pervasive security mechanisms.

3.2.1. Specific Security Mechanisms. The OSI security architecture enumerates the following eight specific security mechanisms:

- *Encipherment*, which refers to the application of cryptographic techniques to encrypt data and to transform it in a form that is not intelligible by an outsider (i.e., somebody not knowing a particular cryptographic key). As such, encipherment can be directly used to protect the confidentiality of data units and traffic flow information or indirectly to support or complement other security mechanisms. Many algorithms and standards can be used for encipherment. Examples include secret key cryptosystems, such as the Data Encryption Standard (DES) and the Advanced Encryption Standard (AES), and public key cryptosystems, such as RSA and ElGamal.

- *Digital signature mechanisms*, which can be used to provide an electronic analog of handwritten signatures for electronic documents. Like handwritten signatures, digital signatures must not be forgeable; a recipient must be able to verify it, and the signer must not be able to repudiate it later. But unlike handwritten signatures, digital signatures incorporate the data (or the hash of the data) that are signed. Different data therefore result in different signatures even if the signatory is unchanged. As of this writing, many countries have or are about the put in place laws for electronic or digital signatures (e.g., the U.S. Electronic Signatures in Global and National Commerce Act). In addition, there are many algorithms and standards that can be used for digital signatures. Examples include RSA, ElGamal, and the Digital Signature Standard (DSS).
- *Access control mechanisms*, which use the authenticated identities of principals, information about these principals, or capabilities to determine and enforce access rights and privileges. If a principal attempts to use an unauthorized resource, or an authorized resource with an improper type of access, the access control function (e.g., the reference monitor) must reject the attempt and may additionally report the incident for the purposes of generating an alarm and recording it as part of a security audit trail. Access control mechanisms and the distinction between discretionary access control (DAC) and mandatory access control (MAC) have been extensively discussed in the computer security literature. They are usually described in terms of subjects, objects, and access rights. A subject is an entity that can access objects. It can be a host, a user, or an application. An object is a resource to which access should be controlled and can range from a single data field in a file to a large program. Access rights specify the level of authority for a subject to access an object, so access rights are defined for each subject/object pair. Examples of UNIX access rights include read, write, and execute. More recently, the idea of role-based access controls (RBACs) has been proposed and adopted by operating system and application software developers.
- *Data integrity mechanisms*, which are used to protect the integrity of either single data units and fields within these data units or sequences of data units and fields within these sequences. Note that data integrity mechanisms, in general, do not protect against replay attacks that work by recording and replaying previously sent messages. Also, protecting the integrity of a sequence of data units and fields within these data units generally requires some form of explicit ordering, such as sequence numbering, timestamping, or cryptographic chaining.
- *Authentication exchange mechanisms*, which are used to verify the claimed identities of principals. In accordance with ITU-T recommendation X.509, the term "strong," is used to refer to an authentication exchange mechanism that uses cryptographic techniques to protect the messages that are exchanged, whereas the term "weak" is used to refer to an authentication exchange mechanism that does not do so. In general, weak authentication exchange mechanisms are vulnerable to passive wiretapping and replay attacks, and the widespread use of weak authentication exchange mechanisms is the single most important vulnerability of contemporary computer networks and distributed systems.
- *Traffic padding mechanisms*, which are used to protect against traffic analysis attacks. Traffic padding refers to the generation of spurious instances of communication, spurious data units, and spurious data within data units. The aim is not to reveal if data that are being transmitted actually represent and encode information. Consequently, traffic padding mechanisms can only be effective if they are protected by some sort of a data confidentiality service. Furthermore, traffic padding is effective in leased lines or circuit-switched networks. It is not particularly useful in packet-switched data networks, such as TCP/IP networks and the Internet.
- *Routing control mechanisms*, which can be used to choose either dynamically or by prearrangement specific routes for data transmission. Communicating systems may, on detection of persistent passive or active attacks, wish to instruct the network service provider to establish a connection via a different route. Similarly, data carrying certain security labels may be forbidden by a security policy to pass through certain networks or links.
- Notarization mechanisms, which can be used to assure certain properties of the data communicated between two or more entities, such as its integrity, origin, time, or destination. The assurance is provided by a trusted third party (TTP) in a testifiable manner.

There are many products that implement specific security mechanisms to provide one (or several) security service(s).

3.2.2. Pervasive Security Mechanisms. Pervasive security mechanisms are not specific to any particular security service and are in general directly related to the level of security required. Some of these mechanisms can also be regarded as aspects of security management. The OSI security architecture enumerates the following five pervasive security mechanisms.

- The general concept of *trusted functionality* can be used to either extend the scope or to establish the effectiveness of other security mechanisms. Any functionality that directly provides, or provides access to, security mechanisms should be trustworthy.
- System resources may have *security labels* associated with them, for example, to indicate sensitivity levels. It is often necessary to convey the appropriate security label with data in transit. A security label may be additional data associated with the data transferred or may be implicit (e.g., implied by the use of a specific key to encipher data or implied by the context of the data such as the source address or route).

- Security-relevant *event detection* can be used to detect apparent violations of security.
- A *security audit* refers to an independent review and examination of system records and activities to test for adequacy of system controls, to ensure compliance with established policy and operational procedures, to detect breaches in security, and to recommend any indicated changes in control, policy, and procedures. Consequently, a *security audit trail* refers to data collected and potentially used to facilitate a security audit.
- *Security recovery* deals with requests from mechanisms such as event handling and management functions, and takes recovery actions as the result of applying a set of rules.

The OSI security architecture can be used to discuss the security properties of any computer network. In the following section, it is used to discuss Internet security as a case study. Similar discussions could be held for any other networking technology, such as wireless networks (e.g., GSM, GPRS, and UMTS networks).

4. CASE STUDY: INTERNET SECURITY

Today, the Internet is omnipresent and issues related to network security are best illustrated using the Internet as a working example for an international information infrastructure. In the past, we have seen many network-based attacks, such as password sniffing, IP spoofing and sequence number guessing, session hijacking, flooding, and other distributed denial of service attacks, as well as exploitations of well-known design limitations and software bugs. In addition, the use and wide deployment of executable content, such as that provided by Java applets and ActiveX controls, for example, have provided new possibilities to attack hosts and entire sites.

There are basically three areas related to Internet security: access control, communication security, and intrusion detection and response. These areas are reviewed next.

4.1. Access Control

In days of old, brick walls were built between buildings in apartment complexes so that if a fire broke out, it would not spread from one building to another. Quite naturally, these walls were called *firewalls*. Today, when a private TCP/IP network (i.e., a corporate intranet) is connected to a public TCP/IP network (e.g., the Internet), its users are usually enabled to communicate with the outside world. At the same time, however, the outside world can interact with the private network and its computer systems. In this situation, an intermediate system can be plugged between the private network and the public network to establish a controlled link, and to erect a security wall or perimeter. The aim of the intermediate system is to protect the private network from network-based attacks that originate from the outside world, and to provide a single choke point where security (i.e., access control) and audit can be imposed. Note that all traffic in and out of the private network can be enforced to pass through this single, narrow choke point. Also note that this point provides a good place to collect information about system and network use and misuse. As a single point of access, the intermediate system can record what occurs between the private network and the outside world. Quite intuitively, these intermediate systems are called *firewall systems*, or *firewalls* in short.

In essence, a firewall system represents a blockade between a privately owned and protected network, which is assumed to be secure and trusted, and another network, typically a public network or the Internet, which is assumed to be nonsecure and untrusted. The purpose of the firewall is to prevent unwanted and unauthorized communications into or out of the protected network. Therefore, it is necessary to define what the terms "unwanted" and "unauthorized" actually mean. This is a policy issue and the importance of an explicitly specified network security or firewall policy is not readily understood today.

In addition to the physical firewall analogy mentioned above, there are many other analogies that may help to better understand and motivate for the use of firewalls. Examples include the tollbooth on a bridge, the ticket booth at a movie theater, the checkout line at a supermarket, the border of a country, and the fact that apartments are usually locked at the entrance and not necessarily at each door. These analogies illustrate the fact that it sometimes makes a lot of sense to aggregate security functions at a single point. A firewall is conceptually similar to locking the doors of a house or employing a doorperson. The objective is to ensure that only properly authenticated and authorized people are able to physically enter the house. Unfortunately, this protection is not foolproof and can be defeated with enough effort. The basic idea is to make the effort too big for an average burglar, causing the burglar to eventually go away and find another, typically more vulnerable, house. However, just in case the burglar does not go away and somehow manages to enter the house, we usually lock up our valuable goods in a safe. According to this analogy, the use of a firewall may not always be sufficient, especially in high-security environments in which we live these days.

Roughly speaking, a firewall is a collection of hardware, software, and policy that is placed between two networks to control data traffic from one network to the other (and vice versa). There are several technologies that can be used to build firewall systems. Examples include (static or dynamic) packet filters, circuit-level gateways (e.g., SOCKS servers), and application-level gateways (i.e., proxy servers). These technologies are usually combined in either a dual-homed firewall or screened subnet firewall configuration with one or several demilitarized zones (DMZs). You may refer to Ref. 2 for an overview and discussion of firewall configurations. In either case, a network security or firewall policy must specify what protocols and services are authorized to traverse the firewall. Typically, a firewall implements a policy that does not restrict outbound connections, but that requires inbound connection to be strongly authenticated by a corresponding application-level gateway. Strong authentication can

be based on one-time password systems, such as SecurID or S/Key, or challenge–response mechanisms.

Today, the market for firewalls is mature and the corresponding products start to differentiate themselves through the provision of additional functionality, such as network address translation (NAT), content screening (e.g., virus scanning), virtual private networking, and intrusion detection. As such, firewalls are likely to stay in corporate environments to provide basic access control and complementary security services to intranet systems.

4.2. Communication Security

According to Shirey [1], the term *communication security* refers to "measures that implement and assure security services in a communication system, particularly those that provide data confidentiality and data integrity and that authenticate communicating entities." The term is usually understood to include cryptographic algorithms and key management methods and processes, devices that implement them, and the lifecycle management of keying material and devices. For all practical purposes, key management is the Achilles heel of any communication security system, and it is certainly the point where an attacker would start with.

Several (cryptographic) security protocols have been developed, proposed, implemented, and partly deployed on the Internet:

- On the *network access layer*, several layer 2 tunneling protocols are in use. Examples include the Point-to-Point Tunneling Protocol (PPTP) and the Layer 2 Tunneling Protocol (L2TP). L2TP is often secured with IPsec as discussed next.
- On the *Internet layer*, several layer 3 tunneling protocols are in use. Most importantly, the IETF (IP security) IPSEC WG has developed and standardized an IPsec protocol with a corresponding key management protocol called Internet Key Exchange (IKE). The IPsec and IKE protocols are in widespread use for virtual private networking. Most firewalls implement the protocols to interconnect network segments or mobile systems.
- On the *transport layer*, several (cryptographic) security protocols layered on top of TCP are in widespread use. Examples include the Secure Sockets Layer (SSL) and the Transport Layer Security (TLS) protocols. These protocols are sometimes also referred to as "session layer security protocols." Unfortunately, there is currently no widely deployed transport layer security protocol layered on top of UDP. This is unfortunate, because SSL/TLS does not provide a solution for UDP-based applications and application protocols.
- On the *application layer*, there are a number of (cryptographic) security protocols that are either integrated into specific applications and application protocols or provide a standardized application programming interface (API). The first class leads to security-enhanced application protocols, such as secure Telnet and secure FTP, whereas the second class leads to authentication and key distribution systems, such as Kerberos. With its use in UNIX and Microsoft operating systems (e.g., Windows 2000 and Windows XP), Kerberos is the most widely deployed authentication and key distribution system in use today.

Above the application layer, there are a few (cryptographic) security protocols that can be used to cryptographically protect messages before they are actually transmitted in computer networks. Examples include Pretty Good Privacy (PGP) and Secure MIME (S/MIME). Another possibility is to use the eXtended Markup Language (XML) with its security features that are currently being standardized by the World Wide Web Consortium (W3C).

Given this variety of (cryptographic) security protocols for the various layers in the Internet model, one may ask what protocol is best or which layer that is best suited to provide security services for Internet applications and users. Unfortunately, both questions are difficult to address and may require different answers for different security services. For example, data confidentiality services can be provided at lower layers, whereas nonrepudiation services are more likely to be provided at higher layers. In either case, the end-to-end argument applies [6]. Roughly speaking, the end-to-end argument states that the function in question (e.g., a security function) can completely and correctly be implemented only with the knowledge of the application standing at the endpoints of the communications system. Therefore, providing that function as a feature of the communications system itself is not possible (sometimes an incomplete version of the function provided by the communications system may be useful as a performance enhancement). This argument should always be kept in mind when network providers argue that security functions can easily be outsourced.

4.3. Intrusion Detection and Response

An *intrusion* refers to a sequence of related actions by a malicious adversary that results in the occurrence of unauthorized security threats to a target computing or networking domain. Similarly, the term *intrusion detection* refers to the process of identifying and responding to intrusions. This process is not an easy one. Nevertheless, there is an increasingly large number of tools that can be used to automate intrusion detection. The tools are commonly referred to as *intrusion detection systems* (IDSs). Although the research community has been actively designing, developing, and testing IDSs for more than a decade, corresponding products have received wider commercial interest only relatively recently. Furthermore, the IETF has chartered an Intrusion Detection Exchange Format (IDWG) WG "to define data formats and exchange procedures for sharing information of interest to intrusion detection and response systems, and to management systems which may need to interact with them."

There are basically two technologies that can be used to implement IDSs: attack signature recognition and anomaly detection.

1. Using *attack signature recognition*, an IDS uses a database with known attack patterns (also known as attack signatures) and an engine that uses this database to detect and recognize attacks. The database can either be local or remote. In either case, the quality of the IDS is as good as the database and its attack patterns as well as the engine that makes use of this database. The situation is similar and quite comparable to the antivirus software (i.e., the database must be updated on a regular basis).
2. Using *anomaly detection*, an IDS uses a database with a formal representation of "normal" (or "normal-looking") user activities and an engine that makes use of this database to detect and recognize attacks. For example, if a user almost always starts up his/her email user agent after having successfully logged onto a system, the IDS's engine may get suspicious if this user starts a Telnet session to a trusted host first. The reason for this activity may be an attacker misusing the account to gain illegitimate access to a remote system. Again, the database can either be local or remote, and the quality of the IDS is as good as the database and its statistical material.

Obviously, it is possible and useful to combine both technologies in a single IDS. The design of IDSs is a new and very active area of research and development. Many technologies that had originally been developed under the umbrella of artificial intelligence (AI) are being reused and applied to the problem of how to reliably detect intrusions. Exemplary technologies include knowledge-based systems, expert systems, neural networks, and fuzzy logic. In fact, some of these technologies have experienced a revival in the field of intrusion detection.

Once an intrusion is detected, it is important to respond in appropriate ways. Therefore, large organizations usually establish and maintain an incidence response team (IRT). For smaller organizations, it is usually more efficient to outsource this task to a commercially operating IRT.

5. CONCLUSIONS AND OUTLOOK

Network security is a hot topic today. Like many other topics related to IT security, network security has many aspects and the OSI security architecture may serve as a primary reference to structure them. In this article, we introduced the OSI security architecture and elaborated on the current state-of-the-art and future perspectives of network security in general, and Internet security in particular. More specifically, we looked into three areas that are particularly important for Internet security: access control, communication security, and intrusion detection and response.

The network security industry has shifted away from an industry that focused primarily on the use of preventive security technologies to an industry that also takes into account the importance of detective and reactive security technologies (under the term "detection and response"). The major argument for this paradigm shift is the insight that preventive security technologies are not complete in the sense that they will always leave vulnerabilities that are still exploitable by attackers, and that administrators must know how to detect attacks and counteract on them. Consequently, detection and response is equally important to prevention and an increasingly large number of companies are providing (security) monitoring services to their customers [7]. The importance of detection and response is likely to continue in the future and we will see many companies starting to specialize in this particular field.

BIOGRAPHY

Rolf Oppliger studied computer science, mathematics, and economics at the University of Berne, Switzerland, where he received M.Sc. and Ph.D. degrees in computer science in 1991 and 1993, respectively. In 1999, he received the Venia legendi for computer science from the University of Zürich, Switzerland. The focus of his professional activities is information technology (IT) security in general, and network security in particular. He has authored nine books, including, for example, the second editions of *Internet and Intranet Security* (Artech House, 2002) and *Security Technologies for the World Wide Web* (Artech House, 2003), frequently speaks at security-related conferences, and regularly publishes papers and articles in scientific magazines and journals. He's the founder and owner of eSECURITY Technologies Rolf Oppliger (www.esecurity.ch), Gümligen, Switzerland works for the Swiss Federal Strategy Unit, Bern, Switzerland, for Information Technology (FSUIT), teaches at the University of Zürich, and serves as editor for the Artech House Computer Security Series and the Swiss digma magazine for data law and information security. He's a member of the Association for Computing Machinery (ACM), the IEEE Computer Society, and served as vice chair of the IFIP TC 11 working group on network security.

BIBLIOGRAPHY

1. R. Shirey, *Internet Security Glossary*, RFC 2828, May 2000.
2. R. Oppliger, *Internet and Intranet Security*, 2nd ed., Artech House, Norwood, MA, 2001.
3. ISO/IEC 7498-2, *Information Processing Systems—Open Systems Interconnection Reference Model—Part 2: Security Architecture*, 1989.
4. ITU X.800, *Security Architecture for Open Systems Interconnection for CCITT Applications*, 1991.
5. J. Zhou, *Non-repudiation in Electronic Commerce*, Artech House, Norwood, MA, 2001.
6. J. H. Saltzer, D. P. Reed, and D. D. Clark, End-to-end arguments in system design, *ACM Trans. Comput. Syst.* **2**(4): 277–288 (1984).
7. B. Schneier, *Secrets and Lies: Digital Security in a Networked World*, Wiley, New York, 2000.

NETWORK TRAFFIC MANAGEMENT

SONIA FAHMY
Purdue University
West Lafayette, Indiana

1. INTRODUCTION

Communication networks have experienced tremendous growth in size, complexity, and heterogeneity since the late 1980s. With the surging popularity of many diverse Internet applications and the increase in content distribution, a "tragedy of the commons" situation has arisen where access must be controlled and congestion must be avoided. Internet traffic can be classified according to the application generating it. A representative list of applications includes video, voice, image, and data in conversational, messaging, distribution, and retrieval modes. These applications are either inelastic (real time), which require end-to-end delay bounds, or elastic, which can wait for data to arrive. Real-time applications can be further subdivided into those that are intolerant to delay, and those that are more tolerant, called *delay-adaptive*. This chapter surveys the required network traffic management building blocks for both types of application traffic. We begin by discussing traffic management objectives and components, and then devote a section for each of these components. We also include a number of case studies that illustrate how these traffic management components can be used to compose services for Internet applications.

1.1. Traffic Management Objectives

Traffic management aims at delivering a negotiated quality of service (QoS) to applications and at controlling congestion. This implies that critical or real-time application traffic may be given better service at network nodes than less critical traffic. In addition, congestion must be controlled to avoid the performance degradation and congestion collapse that occur when network buffers overflow and packets are lost. The network load should not increase beyond a certain optimal operating point, commonly known as the "knee" of the delay throughput curves. This is the point beyond which increasing the load level on the network results in a dramatic increase in end-to-end delay, caused by network congestion and retransmissions. Therefore, the objectives of network traffic management include

1. *Fairness*. Traffic sources should be treated according to some fairness criteria, such as (weighted) max–min fairness (with or without minimum guarantees) [8,42,48], or proportional fairness, which can be tied to pricing through appropriate utility functions [21,52]. *Max–min fairness* gives equal (or weighted) shares to sources sharing a common bottleneck. This means that the share of the constrained (min) sources is maximized, and excess resources are distributed equally among unconstrained sources. Given a configuration with n contending sources, suppose that the ith source is allocated a bandwidth x_i. The allocation vector $\{x_1, x_2, \ldots, x_n\}$ is feasible if all link load levels are less than or equal to 100%. Given an allocation vector, the source with the smallest allocation is, in some sense, the "unhappiest source." We find the feasible vectors that give the maximum allocation to this unhappiest source (thus maximizing the minimum source, or max–min). Then, we remove this "unhappiest source" and reduce the problem to that of the remaining $n - 1$ sources operating on a network with reduced link capacities. We repeat this process until all sources have been allocated the maximum that they can obtain.
2. *Efficient Resource Utilization*. The available resources, such as network buffers, network link bandwidths, processing capabilities, proxy servers, should be efficiently utilized.
3. *Bounded Queuing Delay*. Queuing delay should be small to guarantee low end-to-end delay according to application QoS requirements, and to ensure buffers do not overflow and cause excessive packet loss. Guarantees made to an application can be either deterministic (given for all packets), or statistical. Statistical guarantees can be made in steady state or over specific intervals of time, for instance, over no more than $x\%$ of time intervals will have more than $y\%$ of the packet delays exceed 5 ms.
4. *Stability*. The transmission rates of the sources should not unnecessarily fluctuate in steady state.
5. *Fast Transient Response*. Traffic sources should react rapidly to changing network conditions, such as sudden congestion. Performance should be acceptable even when there is no steady state. Thus, traffic management operations should be robust.
6. *Simplicity*. "Occam's Razor" dictates that entities are not to be multiplied beyond necessity. Traffic management algorithms should have reasonable time and space complexity. This includes scaling to large numbers of users.

Note that traffic management and congestion avoidance are dynamic problems. Static solutions such as increasing buffer size, bandwidth and processing power [43–46], namely, overprovisioning, do not sufficiently address dynamic application needs, especially when some traffic sources are not well behaved.

1.2. Traffic Management Building Blocks

Network traffic management has witnessed a flurry of research activity, especially since the late 1970s. We will use the terms *microflow*, *connection*, and *session* to denote a data stream identified by fields in the Internet Protocol (IP) and the Transmission Control Protocol (TCP) or User Datagram Protocol (UDP) headers, such as source and destination address, protocol identifier, and source and destination ports. The datastream may be unicast (point-to-point), multicast (point-to-multipoint, or multipoint-to-multipoint) from a sending application to a set of receiving applications. We will use *flow* to denote either a microflow or an aggregate of microflows (a macroflow). We will use *end system* to denote a sender,

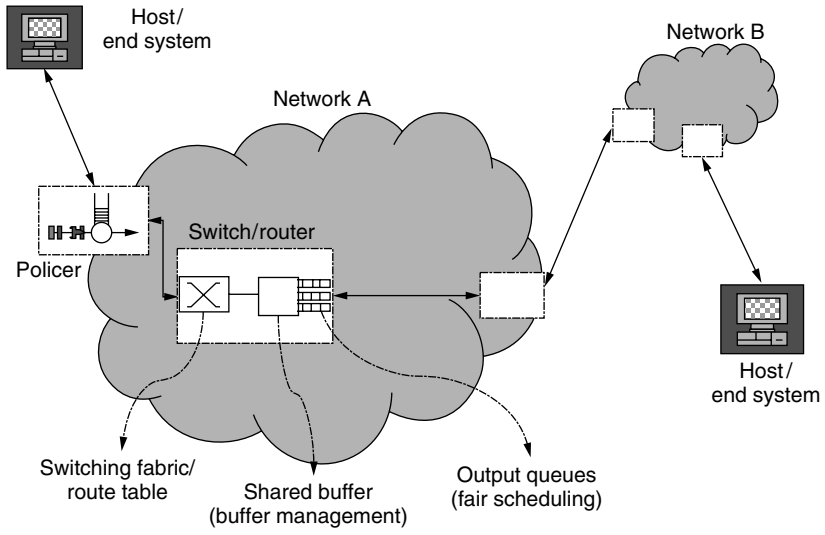


Figure 1. A network can use shaping or policing at the edge, and buffer management and scheduling at core routers and switches to meet guarantees. In combination with constrained routing, admission and policy control, and resource reservation, QoS is provided. Feedback-based congestion control is used to adapt to network conditions.

a receiver, or an edge router. An edge router can be an *ingress* router if it is at the entrance of a domain (with respect to a particular flow or connection); or an *egress* router if it is at the exit of a domain.

Figure 1 illustrates the traffic management components. Constrained routing, connection admission control, policy control, and resource reservations (not shown) are required to ensure that sufficient resources exist for QoS guarantees to be met. Once this is done, traffic shaping and policing, scheduling, and buffer management are required to control resource usage and provide QoS, as shown in the figure. Finally, traffic monitoring and feedback control are important to avoid congestion collapse in computer networks.

An interesting point to note is that the traffic management components bear some similarities to traffic management on transportation highways. For example, an analogy can be drawn between traffic shaping and stoplights that control traffic merging onto a highway on-ramp. Network traffic policing is also analogous to traffic patrol cars that stop speeding vehicles. Packet scheduling resembles several lanes merging onto one lane, or stop signs or lights that control traffic proceeding through an intersection. Buffer management is, in some sense, similar to traffic exiting the highway through various offramps. Finally, constrained routing and congestion control are in some sense similar to listening to traffic reports on your radio and deciding on alternative routes (constrained routing), or deciding not to leave your home yet, if critical highways are congested (congestion control).

The traffic management components in data networks are summarized in Table 1. Admission and policy control, resource reservation and constrained routing are typically performed at a coarse timescale, such as the connection setup time and during renegotiation. Thus we will refer to them as *session-level* (or *connection-level*) operations. Congestion control is invoked as a reaction to network state, and hence the response to network state is on the order of a burst or a round-trip time (time to send a packet from the source to the destination and time to receive feedback back to the source). The remaining traffic management components, namely, traffic shaping, policing, scheduling and buffer management, are performed at the packet level at network switches and routers. Other operations, such as capacity planning and pricing [18,58], operate on timescales on the order of connections, days, or even weeks, and we do not include those in our discussion.

As listed in the fourth column of Table 1, some of the traffic management mechanisms operate at end systems (sources or edge routers) such as traffic shaping; some operate at network switches or routers such as buffer management; and some require both end systems and network switches to cooperate such congestion control using explicit feedback from the network. Congestion control is a closed-loop form of control, while the other operations are typically open-loop, although they may sometimes use measurement to perform better decisions. In the case studies (Section 6), we will see how these building blocks are used in the Internet integrated (IntServ) and differentiated (DiffServ) services, in asynchronous transfer mode (ATM), and with traffic engineering (TE) for label-switched paths.

2. CONSTRAINT-BASED ROUTING

Several routing algorithms that base path selection decisions on policy or quality of service (QoS) have been proposed for the Internet since the mid 1990s. Constraint-based routing usually considers flow aggregates (also known as *macroflows* or *trunks*), rather than individual micro-flows (e.g., a single HTTP connection). Routing constraints may be imposed by administrative policies (Section 2.1), or by the application QoS requirements (Section 2.2).

2.1. Policy Routing

Policy-based routing chooses paths conformant to administrative rules and service agreements. With policy-based routing schemes, the administrators can base routing decisions not only on the destination location but also on factors such as the applications, the protocols used, the size of packets, or the identity of end systems. As the Internet

Table 1. Traffic Management Components and Their Timescales

Timescale	Component	Definition	Location	Open/Closed-Loop
Session level	Admission control	Determines if a new connection/flow requirements can be met without affecting existing connections	Routers/end systems	Open/measured
	Policy control	Determines if a new connection/flow has the administrative permissions to be admitted	Routers/end systems	Open
	Resource Reservation	Sets up resource reservations in network nodes for an admitted connection/flow	Routers/end systems	Open/measured
	Constrained routing	Selects a path based on requirements that are either administrative-oriented (policy-based routing) or service-oriented (QoS routing)	Routers/end systems	Open/measured
Round-trip time (burst level)	Congestion control	Controls the input load to the optimal operating point	End systems with or without router assistance	Closed
Packet level	Traffic shaping	Delays selected packets to smooth bursty traffic	End systems	Open
	Traffic policing	Drops selected packets to conform to a traffic profile	End systems	Open
	Packet scheduling	Determines which packet to transmit next onto the output link	Routers/end systems	Open
	Buffer management	Determines which packets to admit into a buffer	Routers/end systems	Open

continues to grow and diverse Internet services are offered, more stringent administrative constraints can ensure adequate service provisioning and safety from malicious users attempting to obtain services that do not conform to their service agreements or profiles without paying for such services. Policy-based routing can also provide cost savings, load balancing, and basic QoS. Policy constraints are applied before the application of the required QoS constraints (Section 2.2). Policy constraints may be exchanged by the routing protocols while updating route information, or simply provided manually during network configuration. In the latter case, the main problem that may occur is policy rule conflicts.

2.2. QoS Routing

QoS routing can be defined as "a routing mechanism under which paths for flows are determined based on some knowledge of resource availability in the network as well as the QoS requirement of flows" [20]. As most deployed Internet routing strategies are developed for the best-effort model, they are sometimes unsuitable for emerging real-time application requirements. QoS routing extends the best-effort paradigm by finding alternate routes for forwarding flows that cannot be admitted on the shortest existing path. Unlike connectionless best-effort schemes, QoS routing is connection-oriented with resource reservation. QoS routing, however, determines a path only from a source to a destination and does not reserve any resources on that path [80]. A resource reservation technique such as RSVP (Section 6.1) or ATM UNI must then be employed to reserve the required resources. After a path is found and resources are reserved, all packets of the QoS flow must be forwarded through that path. This means that the path must be fixed throughout the lifetime of the flow. This is called "route pinning."

QoS routing dynamically determines feasible paths that also optimize resource usage. Many factors affect the performance of QoS routing solutions, including the particular QoS routing scheme used, the accuracy of information that the QoS routing scheme uses, the network topology, and the network traffic characteristics [63]. A key problem that arises with QoS routing is tractability. Optimizing a path for two or more quality metrics is intractable if the quality metrics are independent and allowed to take real or unbounded integer values [64]. If all metrics except one take bounded integer values, or if all the metrics except one take unbounded integer values but the maximum constraints are bounded, then the problem can be solved in polynomial time [2]. More recent studies show the possibility of performing QoS routing with inaccurate information without suffering significant loss in performance. It was also shown that applying aggregation techniques for scalability does not always negatively impact performance.

3. ADMISSION CONTROL

The Internet protocol (IP) currently supports a best effort service, where no delay or loss guarantees are provided. This service is adequate for non-time-critical applications, or time critical applications under light-load conditions. Under highly overloaded conditions,

however, buffer overflows and queuing delays cause the real-time communication quality to quickly degrade. To support real time applications, a new service model was designed [12]. In this model, both real-time and non-real-time applications share the same infrastructure, thus benefiting from statistical multiplexing gains.

Applications specify their traffic characteristics and their quality of service requirements. Admission control is employed to determine whether these requirements can be met. If they can be met, reservations are made, as discussed in Section 5. Using different classification, policing, shaping, scheduling, and buffer management rules, different applications are serviced with different priorities to ensure that the quality of service requirements are met.

Therefore, admission control is the process where, given the current set of connections and the *traffic characteristics* of a new connection, a decision can be made on whether it is possible to meet the new connection *quality of service requirements*, without jeopardizing the performance of existing connections. Traffic characteristics are commonly described by traffic descriptors. These typically include a subset of the following components: a peak rate, an average rate, and the maximum burst size, that can be enforced by a token bucket or leaky bucket (described in Section 7). For example, in ATM networks the generic cell rate algorithm (GCRA) is used to enforce the peak rate (PCR/CDVT) and average (sustained) rate (SCR/BT) parameters, and the maximum burst size (MBS). QoS requirements are negotiated by the source with the network and used to define the expected quality of service provided by the network. The parameters typically include a maximum delay, delay variation (jitter) and packet loss ratio. For each service, the network guarantees the negotiated QoS parameters if the end system complies with the negotiated traffic contract. For noncompliant traffic, the network need not maintain the QoS objective.

Note that existing connections are accounted for in the admission control algorithm in several possible ways. The declared traffic characteristics of existing connections can be used in the QoS computations. Alternatively, measurement-based admission control (MBAC) can be used, where the new connection declared traffic characteristics are combined with the *measured* traffic characteristics of existing connections, in order to compute whether the QoS would be acceptable to the new connection [49]. This, however, assumes that past measurements are sufficiently correlated with future behavior, which may not always hold. More recently, endpoint admission control has been proposed, where the hosts (the endpoints) probe the network to detect the level of congestion. The host admits a new flow only if the level of congestion is sufficiently low [15].

4. POLICY CONTROL

It is important to firmly control which users are allowed to reserve resources, and how much resources they can reserve. Network managers and service providers must be able to monitor, control, and enforce use of network resources and services based on policies derived from criteria such as the identity of users and applications, traffic or bandwidth requirements, security considerations, or time of day or week. Figure 2 depicts the typical policy control architecture (standardized for the Internet). In this model, a protocol called *Common Open Policy Service* (COPS) is used for policy rule exchange between a policy server [referred to as a *policy decision point* (PDP)] and a network device [referred to as a *policy enforcement point* (PEP)] [11]. The *Lightweight Directory Access Protocol* (LDAP) is used for policy rule retrieval from a policy repository. A policy repository is a server dedicated to the storage and retrieval of policy rules. The Policy Management Console is the coordinator of the entire policy management process.

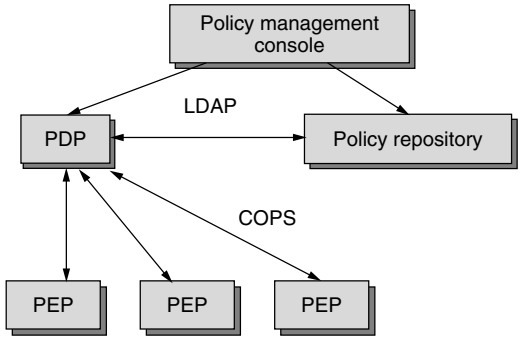


Figure 2. LDAP is used to retrieve the rules from the policy server, and COPS exchanges policy rules among the PDP and PEPs.

5. RESOURCE RESERVATION

A resource reservation protocol is the means by which applications communicate their requirements to the network in an efficient and robust manner. Applications that receive real-time traffic inform the network of their needs, while applications that send real-time traffic inform the receivers and network about their traffic characteristics. The reservation protocol is a "signaling" protocol (a term originating from telephone networks) that installs and maintains reservation state information at each router along the path of a stream. The reservation protocol does not provide any network service; it can be viewed as a "switch state establishment protocol," rather than just a resource reservation protocol. The protocol transfers reservation data as opaque data — it can also transport policy control and traffic control messages.

Resource reservation protocols interact with the *admission control process* to determine whether sufficient resources are available to make the reservation, and the *policy control process* to determine whether the user has permission to make the reservation. If the reservation process gets an acceptance indication from both the admission control and policy control processes, it sends the appropriate parameter values to the packet classifier and packet scheduler. The *packet classifier* determines the QoS class of packets according to the requirements, and the *packet scheduler* (Section 8) and *buffer manager* (Section 9) manage various queues to guarantee the required quality of service. For example, to guarantee the bandwidth and

delay characteristics reserved, a fair packet scheduling scheme can be employed. Fair scheduling isolates datastreams and gives each stream a percentage of the bandwidth on a link. This percentage can be varied by applying weights derived from the reservations [30].

In ATM networks, the User-Network Interface (UNI) protocol establishes resource reservations. In the integrated services framework, the RSVP protocol is used, as discussed in Section 6.1.

6. EXAMPLE ARCHITECTURES

Before we discuss the packet-level and burst-level traffic management components, we will look at four architectures to see how the connection-level building blocks are composed to provide various services.

6.1. Integrated Services and RSVP

An example of a multiservice network is the integrated services framework, which requires resources to be reserved a priori for a given traffic *microflow*. The integrated services framework maps the three application types (delay-intolerant, delay-adaptive, and elastic) onto three service categories: the guaranteed service for delay intolerant applications, the controlled load service for delay adaptive applications, and the currently available best-effort service for elastic applications. The guaranteed service gives firm bounds on the throughput and delay, while the controlled load service tries to approximate the performance of an unloaded packet network [12,81].

Figure 3 illustrates the components of an integrated services router. The Resource Reservation Protocol (RSVP) [13] is the signaling protocol adopted to establish resource reservations state for both unicast and multicast connections. An RSVP sender uses the PATH message to communicate with receiver(s) informing them of microflow characteristics. RSVP provides receiver-initiated reservation of resources, using different reservation *styles* to fit a variety of applications. RSVP receivers periodically alert networks to their interest in a data microflow, using RESV messages that contain the source IP address of the requester and the destination IP address, usually coupled with microflow details. The network then allocates

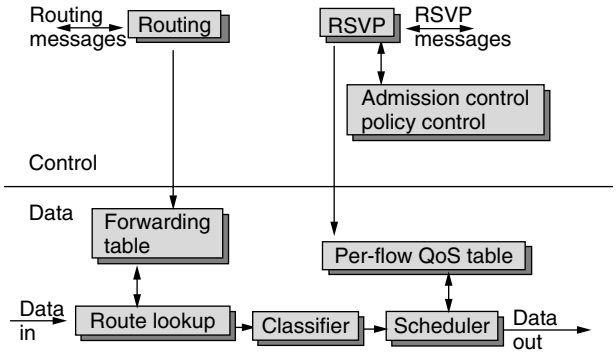


Figure 3. A router with QoS (integrated services) capabilities. RSVP interfaces with admission and policy control, and stores reservation information in a QoS table that is consulted when forwarding a flow.

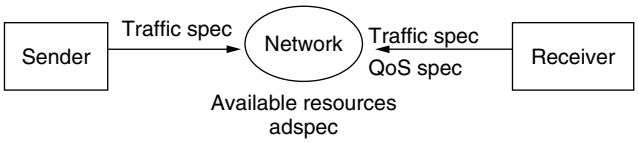


Figure 4. The RSVP sender announces its traffic specifications, and the receiver announces its QoS requirements. The network available resources are checked to see the requested QoS for this traffic can be supported.

the required bandwidth and defines priorities, as shown in Fig. 4. RSVP decouples the packet classification and scheduling from the reservation operation, transporting the messages from the source and destination as opaque data. Periodic renewal of state (soft state) allows networks to be self-correcting despite routing changes and loss of service. This enables routers to understand their current topologies and interfaces, as well as the amount of network bandwidth currently supported.

An RSVP reservation request consists of a **FlowSpec**, specifying the desired QoS, as well as a **FilterSpec**, defining the flow to receive the desired QoS. The FlowSpec is used to set parameters in the packet scheduler, while the FilterSpec is used in the packet classifier. The FlowSpec in a reservation request will generally include a service class and two sets of numeric parameters: (1) an **RSpec** (R for "reserve") that defines the desired QoS, and (2) a **TSpec** (T for "traffic") that describes the data flow. The basic FilterSpec format defined in the present RSVP specification has a very restricted form: sender IP address, and optionally the UDP/TCP source port number.

The main problem with the integrated services model has been its scalability, especially in large public IP networks, which may potentially have millions of concurrent microflows. RSVP exhibits overhead in terms of state, bandwidth, and computation required for each microflow. One of the solutions proposed to this problem is the aggregation of flows, and the simplification of core router state and computation, used in the differentiated services framework.

6.2. Differentiated Services

The differentiated services (DiffServ) framework provides a scalable architecture for Internet service differentiation [9,19]. The main DiffServ design principles are the separation of policy from mechanisms, and pushing complexity to the network domain boundaries, as illustrated in Fig. 5. For a customer to receive DiffServ from its Internet Service Provider (ISP), the customer should have a *service-level agreement* (SLA) agreed on with the ISP. Bandwidth brokers (BBs) [54,62] perform coarse-grained long-term admission and policy control and configure the edge (ingress and egress) routers.

DiffServ core routers are only responsible for forwarding based on the classification performed at the edge. The Differentiated Services Code Point (DSCP) (contained in the IP header DSFIELD/ToS) [61] is used to indicate the forwarding treatment a packet should receive (Fig. 5). DiffServ standardizes a number of per hop behaviors (PHBs) employed in the core routers, including a PHB, expedited

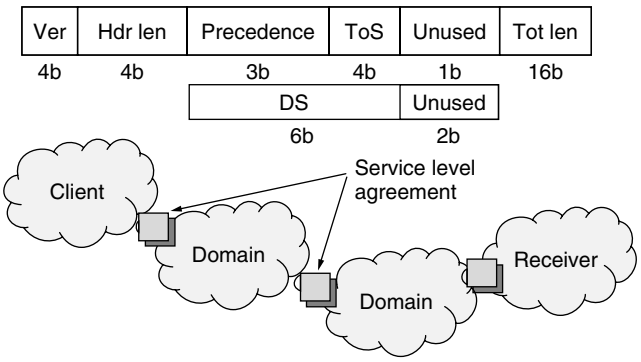


Figure 5. Differentiated Services networks use marking at the edge routers according to bilateral service level agreements, and simple forwarding in the core according to the 6 DS bits in the IP header.

forwarding (EF), and a PHB group, assured forwarding (AF) [37,41]. EF provides a low loss low delay service by employing strict admission control procedures. AF provides a set of better than best effort services for more bursty traffic, by using a number of levels of services, that use multiple queues and multiple drop priorities per queue.

6.3. Asynchronous Transfer Mode (ATM) Networks

Asynchronous transfer mode (ATM) networks were proposed to transport a wide variety of traffic, such as voice, video, and data, in a seamless manner. ATM transports data in fixed size 53 byte-long packets, called *cells*. End systems must set up virtual channel connections (VCCs) of appropriate service categories prior to transmitting information. Service categories are a small number of general ways to provide QoS, which are appropriate for different classes of applications. ATM service categories distinguish real-time from non-real-time services, and provide simple and complex solutions for each case. The added mechanisms in the more complex categories are justified by providing a benefit or economy to a significant subset of the applications [33].

ATM provides six service categories: constant bit rate (CBR), real-time variable bit rate (rt-VBR), non real-time variable bit rate (nrt-VBR), available bit rate (ABR), guaranteed frame rate (GFR), and unspecified bit rate (UBR) [32]. The *constant-bit-rate* (CBR) service category guarantees a constant rate called the *peak cell rate* (PCR). The network guarantees that all cells emitted by the source that conform to this PCR are transferred by the network at PCR. The *real-time variable-bit-rate* (VBR-rt) class is characterized by PCR, sustained cell rate (SCR), and maximum burst size (MBS), which control the bursty nature of traffic. The network attempts to deliver cells of these classes within fixed bounds of cell transfer delay (max-CTD) and cell delay variation (peak-to-peak CDV). *Non-real-time VBR* sources are also specified by PCR, SCR, and MBS, but the network does not specify the CTD and CDV parameters for VBR-nrt.

The *available-bit-rate* (ABR) service category is specified by a PCR as well as a minimum cell rate (MCR), which is guaranteed by the network. Excess bandwidth is shared in a fair manner by the network. We discuss ABR further in Section 10.5. The *unspecified bit rate* (UBR) does not support any service guarantees. UBR VCs are not required to conform to any traffic contract. PCR, however, may be enforced by the network. Switches are not required to perform any congestion control for UBR VCs. When queues become full, switches simply drop cells from UBR connections. Some improvements to UBR, known as UBR+, have been proposed. The *guaranteed-frame-rate* (GFR) service category is an enhancement of UBR that guarantees a minimum rate at the frame level. GFR is different from ABR because it does not use feedback control. The GFR class is intended to be a simple enhancement of UBR that guarantees some minimum rate to application frames.

6.4. Multiprotocol Label Switching

Multiprotocol label switching (MPLS) [73] uses fixed length labels, attached to packets at the ingress router. Forwarding decisions are based entirely on these labels in the interior routers of the MPLS path, as illustrated in Fig. 6. MPLS has made constraint-based routing a viable approach in IP networks [5,6]. Constraint-based routing can reduce manual configuration and intervention required for realization of traffic engineering objectives [6]. The traffic engineer can use administratively configured routes to perform optimizations. This enables a new routing paradigm with special properties, such as being resource-reservation-aware and demand-driven, to be merged with current Internet Gateway routing Protocols (IGPs), such as the Open Shortest Path First (OSPF), or the Intermediate System-Intermediate System (IS-IS) protocols. A constraint-based routing process incorporated in layer 3 and its interaction with MPLS and the current Internet Gateway Protocols (IGPs) is shown in Fig. 7. Constraint-based routing requires schemes for exchanging state information among processes, maintaining this state information, interaction with the current IGP protocols, and accommodating the adaptivity and survivability requirements of MPLS traffic trunks (aggregates of traffic flows) [20].

6.5. Interoperability Among Different Architectures

When QoS networks are deployed, typically only edge networks would be RSVP-enabled, and the core transit network would be DiffServ-enabled, use ATM, or use MPLS with constraint-based routing. In this scenario

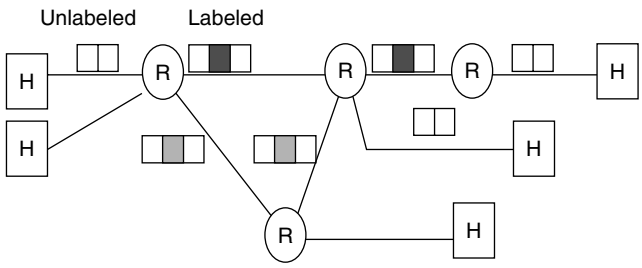


Figure 6. In MPLS, labels are attached to packets and used to perform switching decisions, until the labels are removed. Labels can also be nested.

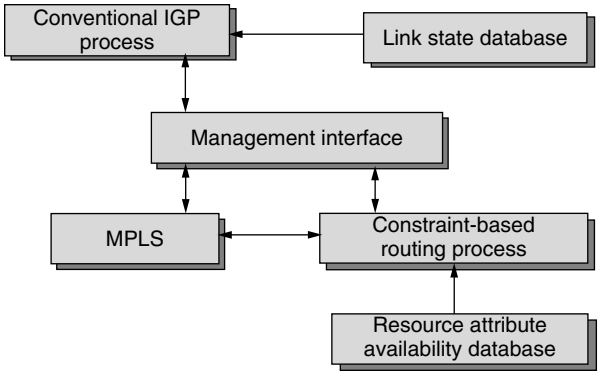


Figure 7. The constraint-based routing process interfaces with MPLS and resource databases.

the RSVP networks (at the edges) may be considered as customers of the transit DiffServ/ATM/MPLS network. The edge routers (at the edge of RSVP and DiffServ networks, for example) would be both RSVP- and DiffServ-capable. RSVP signaling messages are carried transparently through the DiffServ network; only the RSVP-enabled networks process RSVP messages. The DSCP marking can be done either at the host itself or at an intermediate router. RSVP reservations have to be converted into appropriate DiffServ PHBs for achieving end-to-end QoS. MPLS [73] may also be used for establishing label switched paths and trunks based on traffic engineering parameters, as discussed in Section 6.4.

7. POLICING AND SHAPING

In addition to the connection-level operations we have discussed so far, some traffic management operations must be performed for every packet. The end systems and edge routers are responsible for sending data that conforms to a negotiated traffic contract. As previously discussed, a traffic contract typically specifies the average rate, peak rate, and maximum burst size of a traffic flow. An incoming packet is checked against a traffic meter, and a decision is made on whether it is conforming (in profile) or non-conforming (out of profile). The shaping and policing functions (sometimes called usage parameter control (UPC)) have four possible choices when a packet is out of profile:

- *Dropping*. The nonconforming packet can be dropped to ensure that the traffic entering the network conforms to the contract, that is, that traffic is *policed* according to the profile.
- *Marking* (*Tagging*). The nonconforming packet can be marked as a low-priority packet by setting one or more bits in the packet header. In ATM, this is done by setting the value of the cell loss priority (CLP) bit in the ATM header to 1. In DiffServ networks, the DSCP is marked to reflect one of three priority levels, as discussed below. During congestion, the network may choose to discard low-priority packets in preference to high priority packets.
- *Buffering*. The nonconforming packet may be buffered and sent at a later time when it becomes conforming to the contract. This *traffic shaping* function reduces the traffic variation and burstiness and makes traffic smooth. It also provides an upper bound for the rate at which the flow traffic is admitted into the network, thus aiding in computing QoS bounds and buffer requirements [22,23].
- *No Action*. The nonconforming packet may be allowed into the network without any changes. This is typically an undesirable solution because it can cause congestion in the interior of the network.

The leaky-bucket and the token bucket algorithms have been designed for shaping and policing traffic. Partridge [69] describes the *leaky-bucket algorithm*, which was based on ideas discussed by Turner in 1986. The leaky bucket buffers incoming packets in a "bucket" that "leaks" at a certain rate. The algorithm has two input parameters: (1) the depth (size) of the bucket, b, and (2) the rate at which packets are drained out of the bucket, r. The generic cell rate algorithm (GCRA) [32] that is used in ATM is a variation of the leaky bucket.

Whereas the leaky bucket is filled with incoming packets and transmits them (if any are present) at a fixed rate (for shaping), the token bucket indicates whether traffic can be transmitted based on the availability of tokens. Tokens are added to the bucket at a fixed rate r, and can accumulate only up to the bucket depth b (where b controls the maximum burst size). The appropriate number of tokens are consumed when traffic is transmitted. Traffic may be allowed to be sent *in a burst* as long as a sufficient number of tokens is available in the bucket. This is the primary difference between a leaky bucket and a token bucket — a leaky bucket additionally controls the drain rate. Combinations of both are typically used to control the peak rate, average rate, and maximum burst size, according to the service provided.

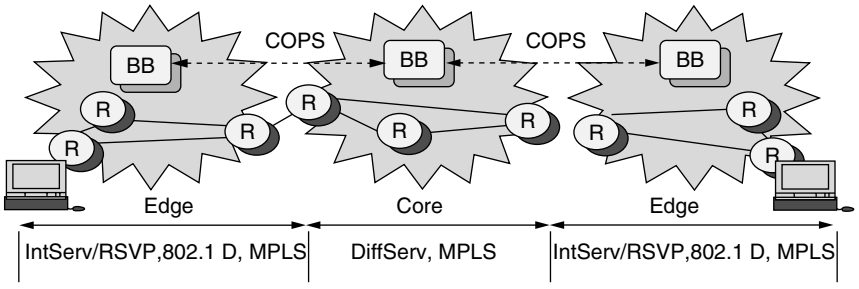


Figure 8. Networks with MPLS label-switched paths, ATM, or QoS (e.g., IntServ/RSVP or DiffServ) capabilities can interoperate. Bandwidth brokers (BBs) use the COPS protocol to exchange policy information among different domains.

For example, in the ATM variable-bit-rate (VBR) service, leaky buckets are used to control the peak rate, PCR, with tolerance CDVT, and to control the sustained (average) rate, SCR, with tolerance BT. The maximum burst size is limited to MBS. In DiffServ networks, the edge router contains meters, markers, droppers, and shapers, collectively referred to as *traffic conditioning functions*. A traffic conditioner may re-mark a traffic stream or discard or shape packets to alter the temporal characteristics of the stream and bring it into compliance with a traffic profile specified by the network administrator. As shown in Fig. 9, incoming traffic passes through a classifier, which is used to select a class for each traffic flow. The meter measures and sorts the classified packets into precedence (priority) levels. The decision (marking, shaping, or dropping) is based on the measurement result.

DiffServ-assured forwarding provides up to three drop precedences for each queue, as depicted in Fig. 10. Assume that the drop precedences are DP0 (green), DP1 (yellow) and DP2 (red), where DP0 means lower precedence to drop, and DP2 means higher (similar to colors in traffic stoplights). A three-color-marker (TCM) is used to mark packets with one of these three precedences. The DSCP is set to one of 3 values according to the DP. Traffic conditioners may also be TCP-aware and choose to protect "critical" TCP packets by marking them as DP0 [35].

8. SCHEDULING

Another packet-level traffic management function is packet scheduling. At every router or switch output port, a *scheduling discipline* must be used decide which packet to transmit next onto the output link (and onto the next hop). Scheduling is important because it resolves contention for a shared resource (the output link) and determines whether real-time applications can be given performance guarantees. Packet scheduling services different queues with different priorities to ensure that the quality of service requirements are met (recall Fig. 10). Multiple queues are needed because if packets from all flows share the same buffer using first-in first-out (FIFO) queuing, it is impossible to isolate packets from various flows. The degree of aggregation (how many microflows share a queue) determine the guarantees that can be made, as well as the state and time complexity of the scheduling discipline.

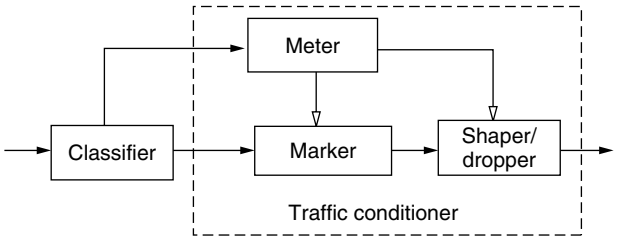


Figure 9. An edge router typically includes a classifier, a meter, a marker, a shaper, and a dropper.

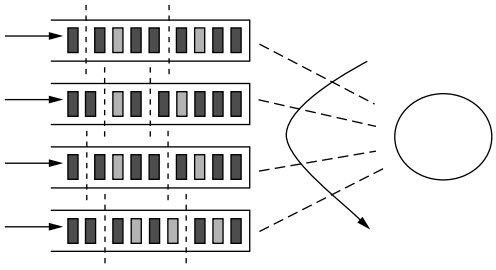


Figure 10. The assured forwarding service queues packets into four queues, with three drop precedences (denoted by three colors) per queue.

Scheduling disciplines aim at (1) meeting performance bounds for guaranteed services, including bandwidth, delay, delay variation (jitter), and loss; (2) providing some level of fairness and protection among flows; (3) being easily integrated with admission control algorithms; and (4) having low state and time complexity. A tradeoff among these goals must be achieved.

Scheduling algorithms may be work-conserving or non-work-conserving. A work-conserving scheduler is idle only when there is no packet awaiting service. Non-work-conserving schedulers may choose to be idle even if they have packets awaiting service, in order to make outgoing traffic more predictable and reduce delay variation (jitter). The best-known work-conserving scheduling discipline is the *generalized processor sharing* (GPS) discipline [67,68]. GPS serves packets as if they are in separate logical queues, servicing an infinitesimally small amount of data from each nonempty queue in turn. This intuitively resembles a *bit-by-bit* round-robin service. Connections can also be assigned weights and can be serviced in proportion to their weights. GPS cannot be directly implemented, but can be emulated as a weighted round robin or deficit round robin [36,75]. Deficit round robin can handle variable packet sizes without having to know the mean packet size of each connection in advance.

Weighted fair queuing (WFQ) approximates *packet-by-packet* GPS [7,24,82]. WFQ computes the time a packet would complete service in a GPS regime, and services packets in the order of these finishing times. A number of variants of WFQ have been developed, including self-clocked fair queuing (SCFQ), virtual clock (VC), and worst-case fair weighted fair queuing (WF^2Q). Other scheduling algorithms include delay and jitter earliest due date (EDD) and stop-and-go. The state overhead of scheduling algorithms can be alleviated if packets carry more information, as in the core-stateless fair queuing approach [77,78].

9. BUFFER MANAGEMENT

In most routers, packets are admitted into the router buffer as long as buffer space is still available. When the buffer is full, incoming packets have to be dropped — which is what is commonly referred to as the "drop tail" policy, since packets are dropped from the tail of the queue. This is the simplest policy to implement. Alternatively, packets may be dropped from the front of the queue, or from random locations within the queue. Such "pushout" mechanisms are typically more expensive.

Partial packet discard (PPD) schemes were first proposed to drop remaining segments, such as ATM cells of

an IP packet, if other segments of the packet have already been dropped. The intuition behind this is that the receiver will anyway discard segments of a packet if the complete packet cannot be reassembled. Therefore, these partial packets should not consume network resources (e.g., bandwidth) on the path between the router where a segment of the packet is discarded, and the receiver. Early packet discard (EPD) [72] has been proposed to extend this notion to drop complete packets when the buffer reaches a certain occupancy, say, 90%, in order to save the remaining capacity for partial segments of admitted packets.

Active queue management (AQM) in routers was later proposed to improve application goodput and response times by detecting congestion *early* and improving fairness among various flows. The main goal of AQM is to drop/mark packets before buffer overflow, in order to (1) give early warning to sources, (2) avoid synchronization among TCP congestion control phases of different flows (TCP congestion control is discussed in Section 10), (3) avoid bias against bursty connections, and (4) punish misbehaving sources.

Active queue management gained significant attention in the early 1990s with the design of the random early detection (RED) algorithm [29]. RED maintains a long-term average of the queue length (buffer occupancy) of a router using a lowpass filter. If this average queue length falls below a certain minimum threshold, all packets are admitted into the queue, as depicted in Fig. 11a. If the average queue length exceeds a certain maximum threshold, all incoming packets are dropped. When the queue length lies between the minimum and maximum thresholds, incoming packets are dropped/marked with a linearly increasing probability up to a maximum drop probability value, p_{max}. RED includes an option known as the "gentle" variant (Fig. 11b). With gentle RED, the packet drop probability varies linearly from p_{max} to 1 as the average queue size varies from th_{max} to twice th_{max}.

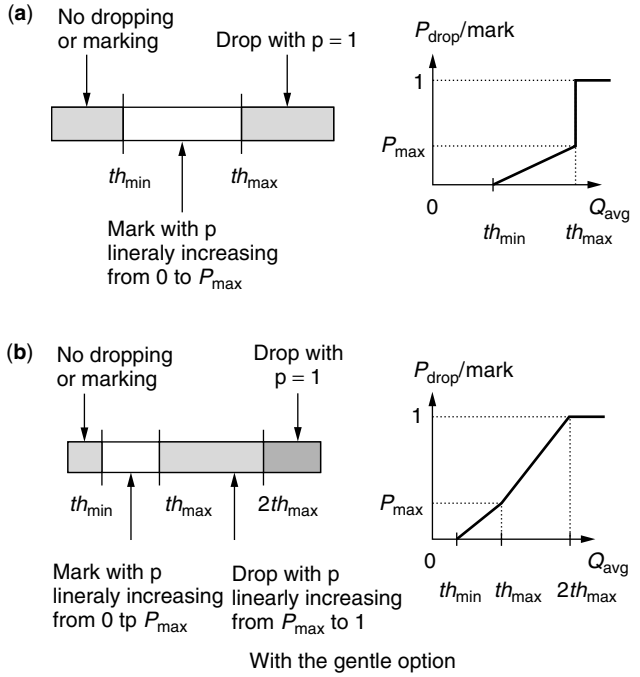


Figure 11. Random early detection (RED) probabilistically drops/marks packets based on average queue length without (**a**) and with (**b**) "gentle" option.

A number of RED variants have appeared, including flow-RED (FRED) [57], stabilized RED (SRED) [65], and BLUE [27]. Although FRED performs best among all RED variants, FRED maintains counts of the buffer occupancies for each flow in order to make better packet admission decisions. This provides the best isolation among flows, especially in the presence of misbehaving flows that send at high rates. However, maintaining per flow packet counts implies that FRED implementation complexity is higher than the other variants. Algorithms for the ATM guaranteed frame rate (GFR) service are also very similar in spirit to FRED. More recently, a number of other algorithms, including random early marking (REM) [4], adaptive virtual queue (AVQ) [55], and the proportional integrator (PI) controller [39], have been proposed. Although RED and these algorithms improve performance over simple drop-tail queues, it is difficult to configure their parameters, and some are complex to implement, so they are still under study.

If the network architecture supports marking packets with different drop precedence values, buffer management algorithms can provide differential drop. For example, in differentiated services networks, within each assured service queue, discrimination among packets can be performed using various mechanisms. The RIO (RED with IN and OUT) algorithm distinguishes between two types of packets, IN and OUT of profile, using two RED instances [19]. Each RED instance is configured with min_{th}, max_{th}, and P_{max} (recall Fig. 11a). Suppose the parameters for the IN profile packets are min_{in}, max_{in}, and P_{max_in}, and for the OUT-of-profile packets are min_{out}, max_{out}, and P_{max_out}. To drop OUT packets earlier than IN packets, min_{out} is chosen to be smaller than min_{in}. The router drops OUT packets more aggressively by setting P_{max_out} higher than P_{max_in}. To realize three drop precedences (red, yellow, and green), three REDs can be used.

10. CONGESTION CONTROL

In addition to connection-level and packet-level traffic management operations, network feedback can be used to control congestion at the burst level. This section discusses how closed-loop feedback operates, using the congestion control algorithms in TCP/IP and ATM networks as case studies.

10.1. TCP Congestion Control

In the year 2000, almost 90% of the Internet traffic used the TCP protocol, although multimedia traffic, especially RTP [74] and RTSP, have been slowly increasing. When congestion collapse was first experienced in the Internet in the 1980s [40,44,48], an investigation resulted into the design of new congestion control algorithms, now an essential part of the TCP protocol. Every TCP connection starts off in the "slow start" phase [40]. The slow-start algorithm uses a variable called *congestion window* (*cwnd*). The sender can only send the minimum of *cwnd* and

the receiver advertised window which we call *rwnd* (for receiver flow control). Slow start tries to reach equilibrium by opening up the window very quickly. The sender initially sets *cwnd* to 1 (or 2) and sending one segment. For each acknowledgment (ACK) the sender receives, *cwnd* is increased by one segment. Increasing by one for every ACK results in exponential increase of *cwnd* over round trips, as shown in Fig. 12. In this sense, the name "slow start" is a misnomer.

TCP uses another variable *ssthresh*, the slow-start threshold, to ensure *cwnd* does not increase exponentially forever. Conceptually, *ssthresh* indicates the "right" window size depending on current network load. The slow-start phase continues as long as *cwnd* is less than *ssthresh*. As soon as *cwnd* crosses *ssthresh*, TCP goes into "congestion avoidance." In the congestion avoidance phase, for each ACK received, *cwnd* is increased by 1/*cwnd* segments. This is *approximately* equivalent to increasing the *cwnd* by one segment in one round trip (an *additive* increase), if every segment (or every other segment) is acknowledged by the destination.

TCP maintains an estimate of the round-trip time (RTT), which is the time it takes for the segment to travel from the sender to the receiver plus the time it takes for the ACK (and/or any data) to travel from the receiver to the sender. The *retransmit timeout* (RTO) maintains the value of the time to wait for an ACK after sending a segment before assuming congestion, timing out and retransmitting the segment. When the TCP sender times out, it assumes the network is congested, and sets *ssthresh* to max(2, min(*cwnd*/2, *rwnd*)) segments, *cwnd* to one, and goes to slow start [1]. The halving of *ssthresh* is a *multiplicative* decrease. The *additive-increase multiplicative-decrease* (AIMD) system has been shown to be stable [17]. This basic version of TCP is called "TCP Tahoe."

10.2. TCP Flavors

Several variations on TCP congestion control have been designed, including TCP Reno [26], New-Reno [38], selective acknowledgments (SACK) [60], and forward acknowledgments (FACK) [59], to recover rapidly from one or multiple segment losses, detected through *duplicate* or *selective* acknowledgments, instead of only through timeouts as with the basic TCP (Tahoe). Other algorithms, such as TCP Vegas [14], use changes in round-trip time estimates, rather than packet loss, to adjust the TCP congestion window. Several TCP variations for wireless networks have also been proposed to operate in environments where bandwidth is limited, bandwidth may be asymmetric, and error rate is high. Because of possibly high error rates, losses are no longer assumed to be due to congestion in these proposals.

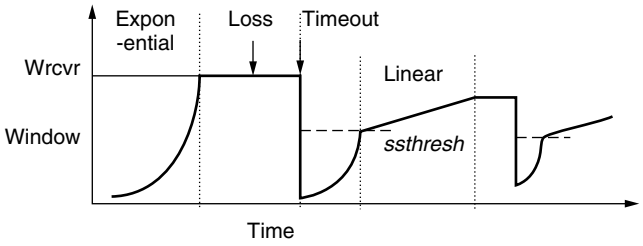


Figure 12. The TCP congestion window grows with ACKs and shrinks when packet loss is detected.

10.3. TCP-Friendly Rate Control

Clever techniques to discover the bottleneck bandwidth and control the source rate accordingly were designed in the early 1990s, including packet pair techniques [53]. Other techniques (proposed for congestion control at the application layer for both unicast and multicast) were used in the Internet Video Service [10], which uses network feedback obtained through the Real Time Control Protocol (RTCP) [74] to control the rate of sources in a video application. More recently, several researchers have investigated how applications can control their transmission *rates* (rather than *windows*) such that it approximates the behavior of TCP. This allows applications running on top of UDP (that do not require reliability) to coexist with TCP connections without starving the TCP connections. Different formulae have been developed that compute the precise "TCP-friendly" application rate. This rate is a function of the connection round trip time, and the frequency of packet loss indications perceived by the connection [66]. Example TCP-friendly protocols include the RAP protocol [71], and TCP-friendly rate control (TFRC) [31].

10.4. Explicit Congestion Indication

As discussed earlier, TCP assumes congestion when it times out waiting for an ACK, or it received duplicate or selective ACKs. This is *implicit feedback* from the network. The *explicit* congestion notification (ECN) option for TCP connections [28,70] allows active queue management mechanisms such as RED to probabilistically mark (rather than drop) packets when the average queue length lies between the two RED thresholds. This is only allowed if both the sender and receiver are ECN-capable (determined at connection setup time). In this case, the receiver echoes back to the sender the fact that some of its packets were marked, so the sender knows that the network is approaching a congested state (Fig. 13). The sender should therefore reduce its congestion window as if the packet was dropped, but need not reduce it drastically as long as it preserves TCP behavior in the long term [56]. The main advantages of ECN are that TCP does not have to wait for a timeout and some packet drops can be avoided.

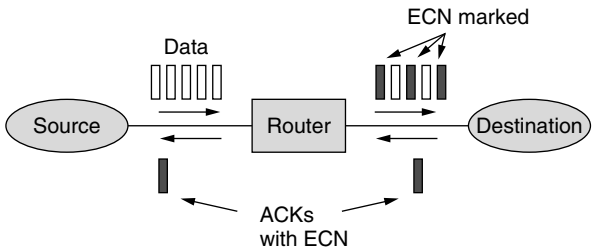


Figure 13. When congestion is incipient, AQM routers mark the ECN bit in TCP packets. The ECN bit is then marked in the ACK packets returning to the sender.

10.5. Explicit Rate Feedback

Instead of using only one bit for explicit feedback, the network can inform the sources of the precise rates they should transmit at. The ATM ABR service (designed before the ECN mechanism was proposed) allows the network to divide the available bandwidth fairly and efficiently among active sources. The ABR traffic management model is (1) "rate-based" because the sources transmit at a specified "rate," rater than using a window; (2) "closed-loop" because, unlike CBR and VBR, there is continuous feedback of control information to the source throughout the connection lifetime; and (3) "end-to-end" because control cells travel from the source to the destination and back to the source [47]. The key attractive features of the ATM ABR service are that it (1) gives sources low cell loss guarantees, (2) minimizes queuing delay, (3) provides possibly nonzero minimum rate guarantees, (4) utilizes bandwidth and buffers efficiently, and (5) gives the contending sources fair shares of the available resources.

The components of the ABR traffic management framework are shown in Fig. 14. To obtain network feedback, the sources send resource management (RM) cells every $Nrm - 1$ (Nrm is a parameter with default value 32) data cells. Destinations simply return these RM cells back to the sources. The RM cells contain the source rate, and several fields that can be used by the network to provide feedback to the sources. These fields are: the explicit rate (ER), the congestion indication (CI) flag and the no increase (NI) flag. The ER field indicates the rate that the network can support for this connection at that particular instant. The ER field is initialized at the source to a rate no greater than the PCR, and the CI and NI flags are usually reset. Each switch on the path *reduces* the ER field to the maximum rate it can support, and sets CI or NI if necessary. When a source receives a returning RM cell, it computes its allowed cell rate (ACR) using its current ACR value, the CI and NI flags, and the ER field of the RM cell [47].

Several algorithms have been developed to compute the ER feedback to be indicated by the network switches to the sources in RM cells [3,16,50,51,76]. The "explicit rate indication for congestion avoidance+" (ERICA+) algorithm [51] computes weighted max–min fair rates (with minimum guarantees) that result in high link utilization and small queuing delay in the network. The algorithm uses the measured load in the forward direction to provide feedback in the reverse direction. The rate is computed as a function of the connection load, the total load, the available capacity, the previous maximum allocation, the normalized weights, and the minimum guarantee. ABR queues can thus be pushed to the edge of a domain [34,79]. Extensions for supporting point-to-multipoint and multipoint-to-point connections are also provided [25].

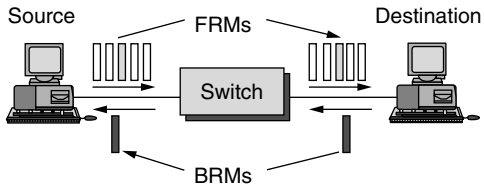


Figure 14. Resource management (RM) cells are sent every Nrm cells in the forward direction, and backward RM cells return to the source with explicit rate information.

BIOGRAPHY

Sonia Fahmy received her PhD degree at the Ohio State University in August 1999. Since then, she has been an Assistant Professor at the Computer Science Department at Purdue University. She has been very active in the Traffic Management working group of the ATM Forum, and has participated in several IETF working groups. Her work is published in over 40 journal and conference papers, and several ATM Forum contributions. She is a member of the ACM, IEEE, Phi Kappa Phi, Sigma Xi, and Upsilon Pi Epsilon, and is listed in the *International Who's Who in Information Technology*. She received the Schlumberger Foundation Technical Merit Award in 2000 and 2001. She has served on the program committees of several conferences, including IEEE INFOCOM, ICNP, and ICC; and co-chaired the SPIE Conference on Scalability and Traffic Control in IP Networks. Her research interests span several areas in the design and evaluation of network architectures and protocols. She is currently investigating multipoint communication, congestion control, and wireless networks. Please see *http://www.cs.purdue.edu/homes/fahmy/* for more information.

BIBLIOGRAPHY

1. M. Allman, V. Paxson, and W. Stevens, *TCP Congestion Control*, RFC 2581, April 1999; *http://www.ietf.org/rfc/rfc2581.txt*; see also *http://tcpsat.lerc.nasa.gov/tcpsat/papers.html.*
2. G. Apostolopoulos, R. Guerin, S. Kamat, and S. K. Tripathi, Quality of service based routing: A performance perspective, *Proc. ACM SIGCOMM*, Sept. 1998, pp. 17–28.
3. A. Arulambalam, X. Chen, and N. Ansari, Allocating fair rates for available bit rate service in ATM networks, *IEEE Commun. Mag.* **34**(11): (Nov. 1996).
4. S. Athuraliya, V. H. Li, S. H. Low, and Q. Yin, REM: Active queue management, *IEEE Network* (May/June 2001) (*http://netlab.caltech.edu/netlab-pub/remaqm.ps.*).
5. D. Awduche et al., *Overview and Principles of Internet Traffic Engineering, Work in Progress*, Aug. 2001; *http://www.ietf.org/*.
6. D. Awduche et al., *Requirements for Traffic Engineering over MPLS*, RFC 2702, Sept. 1999; *http://www.ietf.org/rfc/rfc2702.txt.*
7. J. C. R. Bennett and H. Zhang, Hierarchical packet fair queueing algorithms, *IEEE/ACM Trans. Network.* **5**(5): 675–689 (Oct. 1997).
8. D. Bertsekas and R. Gallagher, *Data Networks*, Prentice-Hall, Englewood Cliffs, NJ, 1992.
9. S. Blake et al., *An Architecture for Differentiated Services*, RFC 2475, Dec. 1998; *http://www.ietf.org/rfc/rfc2475.txt.*

10. J.-C. Bolot, T. Turletti, and I. Wakeman, Scalable feedback control for multicast video distribution in the Internet, *Proc. ACM SIGCOMM*, Sept. 1994.
11. J. Boyle et al., *The COPS* (*Common Open Policy Service*) *Protocol*, RFC 2748; Jan. 2000; *http://www.ietf.org/rfc/rfc2478.txt*.
12. R. Braden, D. Clark, and S. Shenker, *Integrated Services in the Internet Architecture: An Overview*, RFC 1633, June 1994; *http://www.ietf.org/rfc/rfc1633.txt*.
13. R. Braden et al., *Resource ReSerVation Protocol* (*RSVP*), RFC 2205, Sept. 1997; *http://www.ietf.org/rfc/rfc2205.txt*.
14. L. Brakmo, S. O'Malley, and L. Peterson, TCP vegas: New techniques for congestion detection and avoidance, *Proc. ACM SIGCOMM*, Aug. 1994, pp. 24–35; *http://netweb.usc.edu/yaxu/Vegas/Reference/vegas93.ps*.
15. L. Breslau et al., Endpoint admission control: Architectural issues and performance, *Proc. ACM SIGCOMM*, Stockholm, Sweden, Aug. 2000; *http://www.acm.org/sigcomm/sigcomm2000/conf/paper/sigcomm2000-2-2.pdf*.
16. A. Charny, D. Clark, and R. Jain, Congestion control with explicit rate indication, *Proc. ICC'95*, June 1995.
17. D. Chiu and R. Jain, Analysis of the increase/decrease algorithms for congestion avoidance in computer networks, *J. Comput. Networks ISDN Syst.* **17**(1): 1–14 (June 1989) (*http://www.cis.ohio-state.edu/~jain/papers/cong_av.htm*).
18. D. Clark, Internet cost allocation and pricing, in McKnight and Bailey, eds., *Internet Economics*, MIT Press, Cambridge, MA, 1997.
19. D. Clark and W. Fang, Explicit allocation of best effort packet delivery service, *IEEE/ACM Trans. Network.* (Aug. 1998).
20. E. Crawley, R. Nair, B. Rajagopalan, and H. Sandick, *A Framework for QoS-Based Routing in the Internet*, RFC 2386, Aug. 1998; *http://www.ietf.org/rfc/rfc2386.txt*.
21. J. Crowcroft and P. Oechslin, Differentiated end-to-end internet services using a weighted proportional fair sharing tcp, *ACM Comput. Commun. Rev.* **28**(3): (July 1998).
22. R. L. Cruz, A calculus for network delay, Part I: Network elements in isolation, *IEEE Trans. Inform. Theory* **37**(1): 114–131 (Jan. 1991).
23. R. L. Cruz, A calculus for network delay, Part II: Network analysis, *IEEE Trans. Inform. Theory* **37**(1): 132–141 (Jan. 1991).
24. A. Demers, S. Keshav, and S. Shenker, Analysis and simulation of a fair queueing algorithm, *J. Internetwork. Res. Exp.* **1**: 3–26 (1990).
25. S. Fahmy and R. Jain, ABR flow control for multipoint connections, *IEEE Network Mag.* **12**(5): (Sept./Oct. 1998).
26. K. Fall and S. Floyd, Simulation-based comparisons of Tahoe, Reno, and SACK TCP, *ACM Comput. Commun. Rev.* **26**(3): 5–21 (July 1996) (*ftp://ftp.ee.lbl.gov/papers/sacks.ps.Z.*).
27. W. Feng, D. Kandlur, D. Saha, and K. Shin, BLUE: A new class of active queue management algorithms, *Proc. NOSSDAV*, June 2001; also appears as technical report, Univ. Michigan, CSE-TR-387-99, April 1999.
28. S. Floyd, TCP and explicit congestion notification, *ACM Comput. Commun. Rev.* **24**(5): 8–23 (Oct. 1994) (*http://www.aciri.org/floyd/*).
29. S. Floyd and V. Jacobson, Random early detection gateways for congestion avoidance, *IEEE/ACM Trans. Network.* **1**(4): 397–413 (Aug. 1993) (*ftp://ftp.ee.lbl.gov/papers/early.ps.gz*).
30. S. Floyd and V. Jacobson, Link-sharing and resource management models for packet networks, *IEEE/ACM Trans. Network.* **3**(4): (Aug. 1995).
31. S. Floyd, M. Handley, J. Padhye, and J. Widmer, Equation-based congestion control for unicast applications, *Proc. ACM SIGCOMM*, Aug. 2000; multicast extension appears in SIGCOMM 2001.
32. *T*he ATM Forum, *The ATM Forum Traffic Management Specification Version 4.0*; *ftp://ftp.atmforum.com/pub/approved-specs/af-tm-0056.000.ps*, April 1996.
33. M. W. Garrett, Service architecture for ATM: from applications to scheduling, *IEEE Network.* **10**(3): 6–14 (May/June 1996).
34. R. Goyal et al., Per-vc rate allocation techniques for ATM-ABR virtual source virtual destination networks, *Proc. IEEE GLOBECOM* Nov. 1998; *http://www.cis.ohio-state.edu/~jain/papers/globecom98.htm*; see also: S. Kalyanaraman et al., Design considerations for the virtual source/virtual destination (VS/VD) feature in the ABR service of ATM networks, *J. Comput. Networks ISDN Syst.* **30**(19): 1811–1824 (Oct. 1998).
35. A. Habib, S. Fahmy, and B. Bhargava, Design and evaluation of an adaptive traffic conditioner for differentiated services networks, *Proc. IEEE ICCCN* 90–95 (Oct. 2001).
36. E. L. Hahne, Round-robin scheduling for max-min fairness in data networks, *IEEE J. Select. Areas Commun.* **9**(7): 1024–1039 (1991) (*citeseer.nj.nec.com/hahne91roundrobin.html*).
37. J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski, *Assured Forwarding PHB Group*, RFC 2597, June 1999; *http://www.ietf.org/rfc/rfc2597.txt*.
38. J. Hoe, Improving the start-up behavior of a congestion control scheme for TCP, *Proc. ACM SIGCOMM*, 270–280 (Aug. 1996) (*http://www.acm.org/sigcomm/ccr/archive/1996/conf/hoe.ps*).
39. C. V. Hollot, V. Misra, D. Towsley, and W.-B. Gong, On designing improved controllers for AQM routers supporting TCP flows, *Proc. IEEE INFOCOM'2001*, April 2001; *http://www.ieee-infocom.org/2001/*.
40. V. Jacobson, Congestion avoidance and control, *Proc. ACM SIGCOMM* **18**: 314–329, (Aug. 1988) (*ftp://ftp.ee.lbl.gov/papers/congavoid.ps.Z*).
41. V. Jacobson, K. Nichols, and K. Poduri, *An Expedited Forwarding PHB*, RFC 2598, June 1999; *http://www.ietf.org/rfc/rfc2598.txt*.
42. J. M. Jaffe, Bottleneck flow control, *IEEE Trans. Commun.* **COM-29**(7): 954–962 (July 1981).
43. R. Jain, A timeout-based congestion control scheme for window flow-controlled networks, *IEEE J. Select. Areas Commun.* **SAC-4**(7): 1162–1167 (Oct. 1986).
44. R. Jain, A delay-based approach for congestion avoidance in interconnected heterogeneous computer networks, *ACM Comput. Commun. Rev.* **19**(5): 56–71 (Oct. 1989).
45. R. Jain, Congestion control in computer networks: Issues and trends, *IEEE Network Mag.* 24–30 (May 1990).
46. R. Jain, Myths about congestion management in high-speed networks, *Internetwork. Res. Exp.* **3**: 101–113 (1992).

47. R. Jain et al., Source behavior for ATM ABR traffic management: An explanation, *IEEE Commun. Mag.* **34**(11): 50–57 (Nov. 1996) (*http://www.cis.ohio-state.edu/~jain/papers/src_rule.htm*).

48. R. Jain, K. K. Ramakrishnan, and D. M. Chiu, *Congestion Avoidance in Computer Networks with a Connectionless Network Layer*, Digital Equipment Corp., Technical Report DEC-TR-506, Aug. 1987; also in C. Partridge, ed., *Innovations in Internetworking*, Artech House, Norwood, MA, 1988, pp. 140–156.

49. S. Jamin, P. Danzig, S. Shenker, and L. Zhang, A measurement-based admission control algorithm for integrated services packet networks, *Proc. ACM SIGCOMM '95*, 1995, pp. 2–13.

50. L. Kalampoukas, A. Varma, and K. K. Ramakrishnan, An efficient rate allocation algorithm for ATM networks providing max-min fairness, *Proc. 6th IFIP Int. Conf. High Performance Networking*, Sept. 1995.

51. S. Kalyanaraman et al., The ERICA switch algorithm for ABR traffic management in ATM networks, *IEEE/ACM Trans. Network.* **8**(1): 87–98 (Feb. 2000) (*http://www.cis.ohio-state.edu/~jain/papers/erica.htm*).

52. F. Kelly, A. Maulloo, and D. Tan, Rate control in communication networks: Shadow prices, proportional fairness and stability, *J. Oper. Res. Soc.* **49**: 237–252 (1998).

53. S. Keshav, A control-theoretic approach to flow control, *Proc. ACM SIGCOMM '91*, 1991, pp. 3–15.

54. P. Key, Service differentiation: Congestion pricing, brokers and bandwidth futures, *Proc. NOSSDAV*, Basking Ridge, NJ, June 1999; *http://www.nossdav.org/1999/papers/75-1645029201.ps.gz*.

55. S. Kunniyur and R. Srikant, A time-scale decomposition approach to decentralized ECN marking, *Proc. IEEE INFOCOM'2001*, April 2001; *http://www.ieee-infocom.org/2001/*.

56. M. Kwon and S. Fahmy, TCP increase/decrease behavior for explicit congestion notification (ECN), *Proc. IEEE ICC*, April 2002; *http://www.cs.purdue.edu/homes/fahmy/*.

57. D. Lin and R. Morris, Dynamics of random early detection, *Proc. ACM SIGCOMM* **27**: 127–136 (Sept. 1997).

58. J. K. MacKie-Mason and H. R. Varian, *Pricing the Internet*, Dept. Economics, Univ. Michigan, Ann Arbor, 1993.

59. M. Mathis and J. Mahdavi, Forward acknowledgment: Refining TCP congestion control, *Proc. ACM SIGCOMM* (Aug. 1996) (*http://www.psc.edu/networking/papers/papers.html*).

60. M. Mathis, J. Mahdavi, S. Floyd, and A. Romanow, *TCP Selective Acknowledgement Options*, RFC 2018, Oct. 1996; *http://www.ietf.org/rfc/rfc2018.txt*.

61. K. Nichols, S. Blake, F. Baker, and D. Black, *Definition of the Differentiated Service Field (DS Field) in the IPv4 and IPv6 Headers*, RFC 2474, Dec. 1998; *http://www.ietf.org/rfc/rfc2474.txt*.

62. K. Nichols, V. Jacobson, and L. Zhang, *A Two-Bit Differentiated Services Architecture for the Internet*, RFC 2638, July 1999; *http://www.ietf.org/rfc/rfc2638.txt*.

63. A. Orda and A. Sprintson, QoS routing: The precomputation perspective, *Proc. IEEE INFOCOM*, March 2000.

64. A. Orda, Routing with end to end QoS guarantees in broadband networks, *IEEE INFOCOM'98*, April 1998.

65. Teunis J. Ott, T. V. Lakshman, and Larry H. Wong, SRED: Stabilized RED, *Proc. IEEE INFOCOM*, March 1999.

66. J. Padhye, V. Firoiu, D. Towsley, and J. Kurose, Modeling TCP throughput: A simple model and its empirical validation, *Proc. ACM SIGCOMM* **28**: 303–314 (Sept. 1998) (*http://gaia.cs.umass.edu/*).

67. A. Parekh and R. Gallager, A generalized processor sharing approach to flow control in integrated services networks: The single-node case, *IEEE/ACM Trans. Network.* **1**(3): 344–357 (1993).

68. A. Parekh and R. Gallager, A generalized processor sharing approach to flow control in integrated services networks: The multiple node case, *IEEE/ACM Trans. Network.* **2**(2): 137–150 (1994).

69. C. Partridge, *Gigabit Networking*, Addison-Wesley, Reading, MA, 1993.

70. K. Ramakrishnan and S. Floyd, *A proposal to add explicit congestion notification (ECN) to IP*, RFC 2481, Jan. 1999; *http://www.ietf.org/rfc/rfc2481.txt*.

71. R. Rejaie, M. Handley, and D. Estrin, An end-to-end rate-based congestion control mechanism for realtime streams in the internet, *Proc. IEEE INFOCOM*, New York, March 1999; *http://www.ieee-infocom.org/1999/papers/09e_03.pdf*.

72. A. Romanow and S. Floyd, Dynamics of TCP traffic over ATM networks, *IEEE J. Select. Areas Commun.* **13**(4): 633–641 (May 1995).

73. E. Rosen, A. Viswanathan, and R. Callon, *Multiprotocol Label Switching Architecture*, RFC 3031, Jan. 2001; *http://www.ietf.org/rfc/rfc3031.txt*.

74. H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, *RTP: A Transport Protocol for Real-Time Applications*, RFC 1889, 1996; *http://www.ietf.org/rfc/rfc1889.txt*.

75. M. Shreedhar and G. Varghese, Efficient fair queueing using deficit round robin, *Proc. ACM SIGCOMM* 231–243 (Sept. 1995).

76. K. Siu and T. Tzeng, Intelligent congestion control for ABR service in ATM networks, *Comput. Commun. Rev.* **24**(5): 81–106 (Oct. 1995).

77. I. Stoica, S. Shenker, and H. Zhang, Core-stateless fair queueing: A scalable architecture to approximate fair bandwidth allocations in high speed networks, *Proc. ACM SIGCOMM* (Sept. 1998).

78. I. Stoica and H. Zhang, Providing guaranteed service without per flow management, *ACM Comput. Commun. Rev.* **29**(4): 81–94 (Oct. 1999) (*http://redriver.cmcl.cs.cmu.edu/~hzhang-ftp/SIGCOM99.ps.gz*).

79. B. Vandalore et al., QoS and multipoint support for multimedia applications over the ATM ABR service, *IEEE Commun. Mag.* 53–57 (Jan. 1999) (*http://www.cis.ohio-state.edu/~jain/papers/multabr.htm*).

80. Z. Wang and J. Crowcroft, Quality-of-service routing for supporting multimedia applications, *IEEE J. Select. Areas Commun.* **14**(7): 1228–1234 (Sept. 1996).

81. P. P. White, RSVP and integrated services in the Internet: A tutorial, *IEEE Commun. Mag.* (May 1997).

82. H. Zhang, Service disciplines for guaranteed performance service in packet-switching networks, *Proc. IEEE* **83**(10): (Oct. 1995).

NETWORK TRAFFIC MODELING

MICHAEL DEVETSIKIOTIS
North Carolina State University
Raleigh, North Carolina

NELSON L. S. DA FONSECA
Institute of Computing
State University of Campinas
Campinas, Brazil

1. INTRODUCTION

Traffic, that is, data being transmitted, is what telecommunication networks are built to carry. The fascinating and unprecedented "Internet revolution" has led to an ever increasing need for larger amounts of data to be transmitted, as well as fast increasing expectations in terms of the diversity and quality of the transmitted data. Modern networks are expected to accommodate a very heterogeneous traffic mix, including traditional telephone calls, data services, World Wide Web browsing, and video or other multimedia information. In this context, network designers and telecommunication engineers are called on to design, control, and manage networks of increasing transmission speed (bandwidth), size, and complexity. Any effort in network design, control, or management requires decisions and optimization actions that in turn require accurate prediction of the performance of the system under design or control. This is why the science and "art" of traffic modeling has been playing a crucial role in the area of communication network design and operation [7].

The amount of traffic per unit time arriving at a network access point, the number of Internet access requests in an hour or the traffic *workload* through an Internet provider's nodes (routers or switches) is a real physical quantity, even though it consists of bits and bytes and *not* of atoms or molecules. This physical quantity is highly variable with time and space, and appears irregular or, *random*. Furthermore, network traffic usually exhibits visual clusters of activity separated by less active intervals, what is described in the telecommunications lingo, as *bursty* behavior. In order to predict the performance of networks carrying this variable and diverse traffic, researchers and telecommunication engineers utilize analysis (closed-form mathematics), numeric approximations, computer simulation, experimentation with real systems (in the laboratory or in the field), and heuristic or ad hoc projections based on past experience. All of these require, to a great or less degree, some representation or abstraction of *real-life* network traffic, that is, traffic "models."

Traffic modeling has a theoretical/analytic aspect, whereby suitable stochastic models are devised in the mathematical sense, and attributed to different types of data sources and network types. Each model has a number of parameters that determine specific aspects such as mean value, higher moments, autocorrelation function, and marginal density. Such models include [1]

- Renewal models
- Markov and semi-Markov processes
- Autoregressive processes (AR, ARMA, and ARIMA)
- Specially invented processes like Transform-Expand-Sample (TES), SRP, DAR and other
- Long-range dependent, self-similar and multifractal processes

There are also key *computational* and statistical aspects to traffic modeling: After deciding on or hypothesizing about a model (or model family) in the abstract, particular values have to be chosen for the parameters of the model. This usually means performing *matching* or *fitting* where parameter values are estimated statistically from the measured traffic data. Depending on the number of parameters involved, the type of model, and the nature of the data, this task may be far from straightforward and quite time-consuming. The moments to be estimated also depend on the type of network and traffic source, and represent an assumption in themselves. Typical traffic sources include

- Voice, very important for its dominant presence in telephone networks
- Video, especially digital, compressed video (e.g., MPEG)
- Data applications such as FTP, TELNET, SMTP, HTTP
- Traffic in local area and campus networks (LAN and MAN)
- Aggregated traffic on network "trunks" over wide-area networks (WANs)

In this article, we present the most common stochastic processes used for traffic modeling, in Section 2. Such processes can be used either to model the aggregate traffic of several sources (flows, connections, calls) on a network link or can be used to model individual sources, such as the stream generated by a phonecall. Models for specific sources are introduced in Section 3. Some special aspects of traffic modeling related to network performance, namely the concepts of *effective bandwidths* and *envelope processes* are discussed briefly in Section 4. Finally, conclusions and some current open and challenging issues are discussed in Section 5.

2. TRAFFIC MODELS

2.1. General Background

Traffic modeling starts usually by a researcher or telecom engineer collecting samples of traffic during a period of time ("traffic traces") from a specific source and/or at a specific point in the network (e.g., access point, router port, or transmission link). Before stochastic modeling is applied, care must be taken to remove *determinism* and identifying the "residual uncertainty" [16] so that what remains to be modeled is truly stochastic in nature, and *stationary* (i.e., does not have fundamental properties that change with time of the day or month). At a second step, the data are analyzed and a stochastic model is proposed so that a realization of the stochastic process matches the

data trace. A theoretical traffic model has to be checked against several data traces before one can be confident of its accuracy. In what follows, we present stochastic processes commonly used to describe traffic streams.

Network traffic can be *simple* or *compound*. Simple traffic corresponds to single arrivals of discrete data entities (e.g., "packets") and is typically described as a *point process* [9], that is, a sequence of arrival instants $T_1, T_2, \ldots, T_n, \ldots$, with $T_0 = 0$. Point processes can be described equivalently by counting processes and interarrival-time processes. A counting process $\{N(t)\}_{t=0}^{\infty}$ is a continuous-time, nonnegative integer-valued stochastic process, where $N(t)$ is the number of traffic arrivals in the interval $(0, t]$. An interarrival time process is a real-valued random sequence $\{A_n\}_{n=1}^{\infty}$, where $A_n = T_n - T_{n-1}$ is the length of the time interval separating the n-th arrival from the previous one.

Compound traffic consists of *batch arrivals*, that is, multiple units possibly arriving simultaneously at an instant T_n. In the case of compound traffic, we also need to know the real-valued random sequence $\{B_n\}_{n=1}^{\infty}$, where B_n is the (random) number of units in the batch.

In some cases, it is more appropriate or convenient to assume that time is *slotted*, which leads to *discrete-time* traffic models. This means that arrivals may take place only at integer times T_n and interarrival periods are also integer-valued. Furthermore, there are cases where the natural structure of the traffic is such that interarrival times are deterministic or periodic, with only the amount of arriving *workload* changing from arrival to arrival (e.g., compressed video "frames", arriving every $\frac{1}{30}$th of a second).

A simple way to represent a stochastic process is to give the moments of the process — particularly the first and the second moments, which are called the mean, variance, and autocovariance functions. The mean function of the process is defined by $\mu_t = E(X_t)$. The variance function of the process is defined by $\sigma_t^2 = E[(x_t - \mu_t)^2]$, and the autocovariance function between X_{t_1} and X_{t_2} is defined by $\gamma(t_1, t_2) = E[(X_{t_1} - \mu_{t_1})(X_{t_2} - \mu_{t_2})]$.

Another topic that is very relevant in traffic modeling is that of traffic "burstiness." Burstiness is present in a traffic process if the interarrival times process $\{A_n\}$ tends to give rise to runs for several short interarrival times followed by relatively long ones. With typical network traffic exhibiting patterns and bursts that coexist over many magnitudes of time scales (from minutes to hours to days) come the notion of timescale invariance. *Timescale* refers to the change or immunity to change of the process structure on scaling of the time axis. A process $\{X_t\}$ can be defined as scaling invariant if for some $\alpha \in [a, b]$ the process is equal in distribution to its scaled version $\{X_{\alpha t}\}$. If a traffic is not scale-invariant then when studying its behavior as time scales increase, it will show that the bursts and random fluctuations degenerate toward a white noise, nonbursty type of traffic.

The marginal distribution of a process $\{X_t\}$ captures the steady-state first-order distribution of X and is considered the primary characteristic in describing network traffic. Assuming that the process is wide-sense stationary (WSS), the marginal distribution becomes invariant to time and is then defined by the one-dimensional probability density function (PDF): $f_X(x) = f_{X_t}(x) = \frac{d}{dx} Pr[X_t \leq x]$. The PDF describes the probability that the data will assume a value within some given range at any instant of time.

The autocorrelation function of a process $\{X_t\}$ captures the second order measurement of the process and it is used as a supplement to the marginal distribution. The autocorrelation function for network traffic describes the general dependence of the values at another time. Assuming the process is WSS, then the autocorrelation between the data values at times t and $t + k$ is defined as follows:

$$\rho(k) = \frac{E[X_t X_{t+k}] - (E[X_t])^2}{E[(X_t - E[X_t])^2]}$$

where k is called the "lag," the difference or distance between timepoints under consideration. If the autocorrelation function $\rho(k)$ of $\{X_k\}$ is equal to zero for all values of $k \neq 0$, then $\{X_k\}$ is of the *renewal* type. Markov and other *short-range dependent* (SRD) models have a correlation structure that is characterized by an *exponential* decay, which leads to $\sum_k \rho(k) < \infty$.

On the other hand, many real traffic traces exhibit *long-range dependence* (LRD) and can be modeled by self-similar and multifractal models later in this article. For these processes, the autocorrelation function decays slowly (say, polynomially instead of exponentially) in a way that makes the autocorrelation nonsummable: $\sum_k \rho(k) \to \infty$ [23].

2.2. Short-Range Dependent Models

2.2.1. Renewal Models. Renewal models have been used for a long time because of their simplicity and tractability. For this type of traffic, the interarrival times are independent and identically distributed (i.i.d.), with an arbitrary distribution. The major modeling drawback of renewal processes is that the autocorrelation function of A_n is *zero* except for lag $n = 0$. Hence, renewal models seldom capture the behavior of high-speed network traffic in an accurate manner.

Within the renewal family, *Poisson* models are the oldest and most widely used, having been historically closely linked to traditional telephony and the work of A. K. Erlang. A Poisson process is a renewal process with *exponentially* distributed interarrival times with rate λ: $P[A_n \leq t] = 1 - e^{-\lambda t}$. It is also a counting process with $P[N(t) = n] = \frac{(\lambda t)^n e^{-\lambda t}}{n!}$, and independent numbers of arrivals in disjoint intervals. Poisson processes are very appealing due to their attractive memoryless and aggregation properties.

2.2.2. Markov Models. Unlike renewal traffic models, Markov and Markov renewal traffic models [1,7] introduce dependence into the random sequence A_n. Consequently, they can potentially capture traffic burstiness, due to nonzero autocorrelations of A_n. Consider a Markov process $M = \{M(t)\}_{t=0}^{\infty}$ with a discrete state space, where M behaves as follows. It stays in state i for an exponentially distributed holding time that depends on i alone; it then jumps to state j with probability p_{ij}, such that the matrix

$P = [p_{ij}]$ is a probability matrix. In a simple Markov traffic model, each jump of the Markov process corresponds to an arrival, so interarrival times are exponentially distributed, and their rate parameter depends on the state from which the jump occurred. Arrivals may be single, a batch of units or a continuous quantity.

Markov-modulated models constitute another important class of traffic models. Let $M = \{M(t)\}_{t=0}^{\infty}$ be a continuous-time Markov process, with state space of $1, 2, \ldots, m$. Now assume that while M is in state k, the probability law of traffic arrivals is completely determined by k. Thus, the probability law for arrivals is *modulated* by the state of M. The modulating process can be more complicated than a Markov process (so the holding times need not be restricted to exponential random variables), but such models are far less analytically tractable.

The most commonly used Markov modulated model is the *Markov modulated Poisson process* (MMPP) model, which combines a modulating (Markov) process with a modulated Poisson process. In this case, while in state k of M, arrivals occur according to a Poisson process of rate k. As a simple example, consider a 2-state MMPP model, where one state is an ON state with a positive Poisson rate, and the other is an OFF state with a rate of zero. Such models have been widely used to model voice traffic sources.

A semi-Markov process is a generalization of Markov processes, that allows the holding time to follow an arbitrary probability distribution. This destroys the Markov property since times are not exponentially distributed, however it allows for more general models of traffic. When values from a semi-Markov chain are generated, the next state is chosen first, followed by a value for the holding time. If the holding times are ignored, then the sequence of states will be a discrete time Markov chain, referred to as an *embedded* Markov chain.

2.2.3. Autoregressive Models. The autoregressive model of order p, AR(p), is a process $\{X_t\}$ whose current value is expressed as a finite linear combination of previous values of the process plus a white-noise process ε_t: $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t$ where ϕ_i are constants and X_{t-i} are past values of the process at time $t - i$. The recursive form of this model makes it a popular modeling candidate as it makes it straightforward to *generate* an autocorrelated traffic sequence, such as variable-bit-rate (VBR) video traffic [1,7]. However, autoregressive models cannot simultaneously match the empirical marginal distribution of arbitrary traffic such as video.

Another model of the same family is the autoregressive moving-average model of order (p, q), denoted by ARMA (p, q): $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$. Because of the larger number of parameters, ARMA models are more flexible than AR models and can be used in more cases. However, estimation of its parameters is more involved.

2.2.4. Transform–Expand–Sample (TES). Transform–expand–sample (TES) models represent another important class of models appropriate for modeling autocorrelated traffic streams. This family of models aims to capture *both* autocorrelation and *marginal distribution* of the empirical traffic trace; in fact, it was historically the first traffic model explicitly devised to accomplish exactly this dual purpose and specifically for network traffic data. TES models capture stationary, correlated time series and also allow one to generate synthetic streams of real-looking traffic streams to drive simulations of networks [7].

TES models include two types of TES processes: TES$^+$ and TES$^-$. TES$^+$ produces sequences with positive autocorrelation at lag 1, while TES$^-$ produces negative autocorrelation at lag 1. The TES$^+$ process is more suitable for modeling network traffic. To define the TES$^+$ process, we first introduce a modulo 1 operation. The modulo 1 of a real number x, denoted by $\langle x \rangle$, is defined as $\langle x \rangle = x - \lfloor x \rfloor$, where $\lfloor x \rfloor$ is the maximum integer less than x. The recursive construction of the background TES$^+$ process is defined by

$$U_n^+ = \begin{cases} U_0^+ & n = 0 \\ \langle U_{n-1}^+ + V_n \rangle & n > 0 \end{cases}$$

where $\{V_n\}$ is a sequence of IID random variables referred to as *innovations* and U_0^+ is uniformly distributed on $[0, 1)$ and independent of $\{V_n\}$. The resulting sequence $\{U_n^+\}$ has a $[0, 1)$ uniform marginal distribution, and autocorrelation function determined by the probability density function $f_V(t)$ of V_n. The choice of $f_V(t)$ determines the correlation structure of the resulting process. From this background sequence the output process of the model referred to as the foreground sequence, $\{X_n^+\}$ is created by "distorting" each U_n^+ by $X_n^+ = F^{-1}(U_n^+)$, where F is the marginal distribution of the empirical data [9].

2.2.5. Other Short-Range Dependent Models. Another interesting model is the *spatial renewal process* (SRP), which efficiently models processes exhibiting arbitrary marginal distribution and aperiodically decaying autocorrelation (see the paper by Taralp et al. [20] and references cited therein).

A *discrete autoregressive model* of order p, denoted as DAR(p), generates a stationary sequence of discrete random variables with an arbitrary probability distribution and with an autocorrelation structure similar to that of an AR(p). DAR(1) is a special case of DAR(p) process; it has a smaller number of parameters than do general Markov chains, simpler parameter estimation, and can match arbitrary distributions. Moreover, the analytic queuing performance is tractable (see paper by Adas [1] and references cited therein).

2.3. Long-Range Dependent and Self-Similar Traffic Models

Measurements and statistical analysis of real traces performed during the 1990s revealed that traffic exhibits large irregularities (*burstiness*) both in terms of extreme variability of traffic intensities as well as persistent autocorrelation. Network traffic often looks extremely irregular at different timescales [12,17], and such extreme behavior is not exhibited by the traditional Poisson traffic, which smoothes out when aggregated at coarser timescales. If traffic were to follow a Poisson or Markov arrival process, it would have a characteristic burst length that would tend to be smoothed by averaging over a long enough timescale. Instead, measurements of real traffic indicate consistently

that significant traffic burstiness is present on a wide range of timescales.

This behavior is reminiscent of and has been modeled according to *self-similar* processes. Self-similar or *fractal* modeling has been used in a number of research areas such as hydrology, financial mathematics, telecommunications, and chaotic dynamics [4,24]. Internet traffic, and more generally broadband network traffic, is an area where fractal modeling has become popular more recently. Such modeling has also been related to the observation of ON–OFF traffic with "heavy-tailed" distribution [23].

2.3.1. Heavy-Tailed ON–OFF Models. The fractal nature of network traffic is consistent with and predicted by the behavior of the individual connections that produce the aggregate traffic stream. In WAN traffic, individual connections correspond to "sessions," where a session starts at a random point in time, generates packets or bytes for some time and then stops transmitting. On the other hand, in LAN traffic, individual connections correspond to an individual source–destination pair. Individual connections are generally described using simple traffic models such as ON–OFF sources.

Traditional ON–OFF models assume finite variance distributions for the duration of the ON and the OFF periods. The aggregation of a large number of such processes results in processes with very small correlations. On the other hand, a positive random variable Y is called "heavy-tailed with tail index α," if it satisfies: $P[Y > y] = 1 - F(y) \approx cy^{-\alpha}, y \to \infty, 0 < \alpha < 2$, where $C > 0$ is a finite constant independent of y. This distribution has infinite variance. Furthermore, if $1 < \alpha < 2$, then it has a finite mean. The superposition of many such sources was shown to produce aggregate traffic that exhibits long-range dependence and even self-similarity [12,22].

2.3.2. Monofractal Models. Self-similarity in a process indicates that some aspect of the process is *invariant* under scale-changing transformations, such as "zooming" in or out. In network traffic, this is observed when traffic becomes bursty, exhibiting significant variability, on many or all timescales. The appeal and modeling convenience of self-similar processes lies in the fact that the degree of self-similarity of a series can be expressed using only one parameter. The *Hurst* parameter, H, describes the speed of decay of the series autocorrelation function. For self-similar series the value of H is between 0.5 and 1. The degree of self-similarity increases as the Hurst parameter approaches unity.

A process $\{X_k\}$ whose autocorrelation function, $\rho(k)$, takes the form $\rho(k) \approx ck^{-\beta}, 0 < \beta < 1$, for large k and a constant $c > 0$, is said to be *long-range dependent*. This implies that the autocorrelation function decays slowly and is not summable, thus $\sum_k \rho(k) \to \infty$. Figure 1 shows the autocorrelation as a function of time for streams with different H values. Note that for streams with greater H the autocorrelation decays more slowly as a function of time.

In the case of traffic traces, self-similarity is used in the distributional sense: when viewed at varying

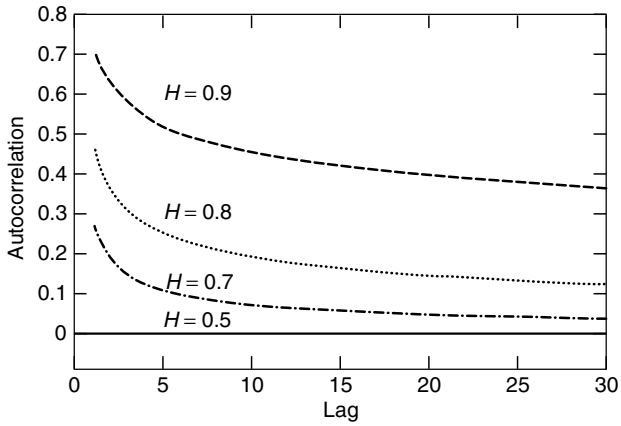


Figure 1. The autocorrelation as a function of time for different values of H.

scales, the object's distribution remains unchanged. Equivalence in distribution between X and Y is denoted by $X \stackrel{d}{=} Y$. We provide in the following certain common definitions of self-similar traffic processes, following Tsybakov and Georganas [22]. Let $X = \{X_t : t = 1, 2, 3, \ldots\}$ be a second-order stationary sequence with mean $\mu = E[X_t]$, variance $\sigma^2 = \text{var}(X_t)$, and autocorrelation function $r(k) = \dfrac{E[(X_{t+k} - \mu)(X_t - \mu)]}{\sigma^2}$. Let $X^{(m)}(t) = \frac{1}{m}(X_{tm-m+1} + \cdots + X_{tm}), m = 1, 2, 3, \ldots$, be the corresponding aggregated sequence with level of aggregation m, obtained by dividing the original sequence X into nonoverlapping blocks of size m and averaging over each block. The index t labels the block. For each $m = 1, 2, 3, \ldots$ let $\mathbf{X}^{(m)} = \{X^{(m)}(k) : k = 1, 2, 3, \ldots\}$ denote the averaged process with autocorrelation function $r^{(m)}(k)$.

- A process X is called *exactly second-order self-similar* with parameter $H = 1 - (\frac{\beta}{2}), 0 < \beta < 1$ if its correlation coefficient is $r(k) = \frac{1}{2}[(k+1)^{2-\beta} - 2k^{2-\beta} + (k-1)^{2-\beta}], k = 1, 2, 3, \ldots$.
- A strict-sense stationary process X is called *strictly self-similar* with parameter $H = 1 - (\frac{\beta}{2}), 0 < \beta < 1$, if $X \stackrel{d}{=} m^{1-H}X^{(m)}$. If X is strictly self-similar, then it is also exactly second-order self-similar. The opposite is not true, except for Gaussian processes.
- A process X is called asymptotically second-order self-similar with parameter $H = 1 - (\frac{\beta}{2}), 0 < \beta < 1$, if $\lim_{m\to\infty} r^{(m)}(k) = \frac{1}{2}\delta^2(k^{2-\beta}), k = 1, 2, 3, \ldots$ where $\delta^2(f(x)) = f(x + \frac{1}{2}) - f(x - \frac{1}{2})$.
- A strict-sense stationary process X is called strictly asymptotically self-similar if $X^{(m)} \stackrel{d}{=} X, m \to \infty$. Note that a strictly asymptotically self-similar process is not necessarily asymptotically second-order self-similar.

2.3.3. Fractal Gaussian Noise and Fractal Brownian Motion. The fractal Brownian motion (FBM) is a self-similar process with Gaussian stationary increments [14]. The increment process is called *fractal Gaussian noise*, and its autocorrelation function is invariant under aggregation and is given by $r(k) = 1/2[|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}]$.

The FBM process accurately models Ethernet, ATM, and FDDI traffic, as well as video sources. The aggregate of ON–OFF sources with heavy tails tends to an FBM.

The analysis of a queuing system with FBM input is quite challenging. However, it becomes manageable if the fractal Brownian traffic [15] process is used instead. The fractal Brownian traffic is defined as the fluid input in time interval $(s, t]$, and is given by $A(s,t) = m(t-s) + \sigma(Z_t - Z_s)$ where m is the mean input rate, σ^2 is the variance of traffic in a given time unit, and Z_x is a normalized fractal Brownian motion, defined as a centered Gaussian process with stationary increments and variance $E[Z_t^2] = t^{2H}$.

2.3.4. Distorted Gaussian. The distorted Gaussian (DGauss) model begins with a Gaussian process with a given autocorrelation structure and maps it into an appropriate marginal distribution. Examples of this popular traffic generation technique include the autoregressive-to-anything process [20] and the self-similar traffic model [8].

Many techniques exist to generate Gaussian time series (Gaussian in the marginal distribution) with a wide range of autocorrelation decay characteristics. A background Gaussian process Z_k is imparted with an autocorrelation structure $\rho'(t)$ and is run through a fitting function $X_k = F_X^{-1}(F_N(Z_k))$ to map its values into an appropriate distribution. Because of the background–foreground transformations, precompensation is applied to the background autocorrelation ρ' such that the resulting output autocorrelation ρ matches the desired specification [8].

2.3.5. Fractal Lévy Motion. Laskin et al. [11] introduced a teletraffic model that takes into account, in addition to the Hurst parameter $H \in [\frac{1}{2}, 1)$, the Lévy parameter $\alpha \in (1,2]$. This was the so-called *fractional Lévy motion* (fLm), mentioned by Mandelbrot [14]. Two important subclasses of Lévy motion exist: (1) the well-known ordinary Lévy motion (oLm), an α-stable process (distributed in the sense of P. Lévy) with independent increments, which is a generalization of the ordinary Brownian motion (the Wiener process); and (2) the fractional Lévy motion, a self-similar and stable distributed process, which generalizes the fractional Brownian motion (fBm), has stationary increments and an infinite "span of interdependence."

Several self-similar stable motions have been proposed for traffic modeling. These processes combine, in a natural way, both scaling behavior and extreme local irregularity.

2.4. Multifractal Models

Historically following self-similar models, researchers have been studying also the possibility of modeling network traffic with *multifractal* processes (see book by Park and Willinger [16] and references cited within). It appears that even though measured network traffic is consistent with asymptotic self-similarity, it also exhibits small timescaling features that differ from those observed over larger time scale. This small timescaling behavior has been related to communication protocol-specific mechanisms and end-to-end congestion control algorithms that operate at those small timescales (less than a few hundred milliseconds). Modeling network traffic with multifractals has the potential of capturing the observed scaling phenomena at large as well as small timescales and thus to naturally extend and improve the original self-similar models of measured traffic.

To quantify the local variations of traffic at a particular point in time t_0, let $Y = \{Y(t), 0 < t < 1\}$ denote the traffic rate process representing the total number of packets or bytes sent over a link in an interval $[t_0, t_0 + t]$. The traffic has a *local scaling component* $\alpha(t_0)$ at time t_0 if the traffic rate process behaves like $t^{\alpha(t_0)}$ as $t \to 0$. In this context, $\alpha(t_0) > 1$ relates to instants with low intensity levels or small local variations, and $\alpha(t_0) < 1$ is found in regions with high level of burstiness or local irregularities.

If $\alpha(t_0)$ is constant for all t_0, then the traffic is *monofractal*. Equivalently, if $\alpha(t_0) = H$ for all t_0, then the traffic is exactly self-similar, with Hurst parameter H. On the other hand, if $\alpha(t_0)$ is not constant and varies with time, the traffic is *multifractal*.

The multifractal appearance of WAN traffic is attributed to the existence of certain multiplicative mechanisms in the background. Multifractal processes are well modeled using multiplicative processes or "conservative cascades." The latter are a fragmentation mechanism, which preserves the mass of the initial set (or does so in the expected value sense). The generator of the cascade is called the fragmentation rule and the mathematical construct that describes the way mass is being redistributed is called the limiting object or multifractal. Modern data networks together with their protocols and controls can be viewed as specifying the mechanisms and rules of a process that fragments units of information at one layer in the networking hierarchy into smaller units at the next layer, and so on.

Multifractal processes are a generalization of self-similar processes. Hence, self-similar processes are also multifractal, but the reverse is not always true. This leads to the important modeling question: Which of the two types of models is more appropriate in a given case? A method for distinguishing between the two models has been proposed [19]. Their conclusion was that traffic traces from environments were well modeled using self-similar models and that more sophisticated models such as multifractals were not needed. On the other hand, in WAN environments, there were cases where self-similar models were not deemed adequate and where multifractal models appeared to be more appropriate.

2.5. Fluid Traffic Models

In fluid traffic modeling, individual units such as packets, are not explicitly modeled. Instead, traffic is viewed as a "stream of fluid" arriving at a certain *rate* that may be changing. Fluid models can simplify analysis due to their lower "resolution" or level of detail. More importantly, fluid models can make network simulation much more efficient, since the computer representation of the fluid traffic requires much fewer "events" (e.g., rate changes) that need to be tracked.

In modern high-speed networks such as asynchronous transfer mode (ATM) networks, the size of individual packets is often fixed and very small (e.g., 53 bytes), relative to the total transmission speed and aggregate volume of

information being transmitted (e.g., hundreds of megabits or gigabits per second). Therefore fluid modeling may be appropriate in such cases and, in general, whenever individual packets can be regarded as effectively insignificant with respect to the total traffic. The validity of this approximation depends heavily on the timescale involved as well as the point of interest inside the network (e.g., access points versus large routers in the middle of the network).

Fluid models [9] typically assume that sources are bursty, commonly of the ON–OFF type. In the OFF state, there is *no* traffic arriving, while in the ON state traffic arrives at a constant rate. To maintain analytic tractability, the durations of ON an OFF periods are assumed exponentially distributed and mutually independent (i.e., they form an alternating renewal process).

3. SOURCE MODELS

In this section, the modeling of different type of traffic sources is discussed. The flow generated by some network sources are regulated by the stack of protocols used in the network. Such type of sources is called *elastic sources*. Sources whose flow do not depend on network protocol are called *streaming sources*. First, streaming multimedia sources are introduced, followed by the modeling of elastic sources. This section concludes with a general characterization of traffic streams, called *effective bandwidth*.

3.1. Data

Datastreams were traditionally modeled by Poisson processes. The rationality behind it was that the superposition of several independent renewal processes tends to a Poisson process.

The nature of traffic changes as new applications becomes a significant part of the network traffic. SMTP, email, TELNET, and FTP were responsible for most of the traffic in pre-Web time. As the use of Web services became predominant, Internet traffic began to present new patterns. Most of today is network traffic is based on the Transmission Control Protocol (TCP). Internet traffic observed at long timescales exhibits self-similarity. However, at long timescales, typically shorter than a round–trip time, Internet traffic presents high variability. At short timescales, Internet traffic marginal distribution is non-Gaussian and the scaling exponent of the variance is smaller than the asymptotic exponent. In other words, at short timescales, Internet traffic exhibits multifractal scaling, with different moments of the traffic showing scaling described by distinct exponents. However, at long timescales it can be modeled as self-similar [4,5]. Such behavior is originated by the complex interaction between network protocols that governs the network flow and TCP sources.

3.2. Voice

The packetstream from a voice source can be characterized by an ON–OFF model; Thus, during silent periods no packet is generated and during "talkspurt" periods, packets are generated either at exponentially distributed intervals or at constant intervals depending whether compression algorithms are used. The residence time in each state is exponentially distributed.

A popular approach to analyze a multiplexer fed by several ON–OFF sources is to use Markov modulated processes to mimic the superposition process. The arrival rates and the transition probabilities of the underlying Markov chain are defined in a way that certain statistics of the Markov modulated process have the same numerical value of the corresponding statistics of the superposition process. The advantage of adopting a 2-state process is to keep the complexity of both the matching procedure and the queuing solution low. In a 2-state MMPP (Fig. 2) there are only four parameters to be determined: the arrival rate and the sojourn time in each state. Several procedures are available to set these four parameters. Most of procedures consider 2 superstates: the underloaded and the overloaded states [18]. In the overload state, the packet generation rate (due to the number of source in state ON)

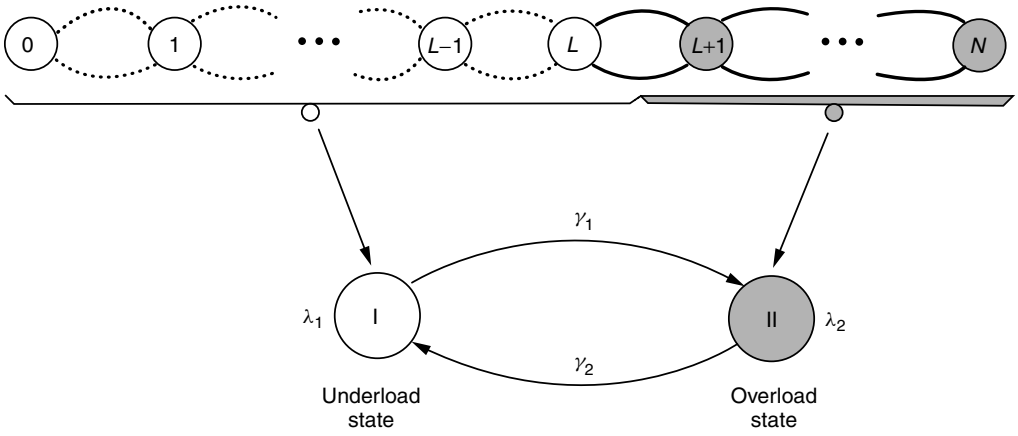


Figure 2. Modeling the superposition of voice sources as a 2-state MMPP. The states of the original Markov chain represent the number of sources in state ON and the arrival rate, in the nth state is n times the arrival rate in state ON. The superstates of the 2-state MMPP correspond to underload overload periods depending on whether the aggregated arrival rate surpasses the channel capacity.

exceeds the server capacity, whereas in the underload state it is below the server capacity.

3.3. Video

The bit rate of a videostream depends not only on the coding algorithm but also on the level of activity of a scene. Whenever there is a scene change, a new scene has to be encoded, generating a high number of bits to be transmitted, and, consequently high bit rates.

The MPEG coding scheme is widely used for several types of applications. MPEG streams consist of a series of frames. In MPEG-2, there are three types of frames: intracoded (I), predictive (P), and bidirectional (B). A periodic sequence of I, B, P frames is called *group of pictures* (GOP). An MPEG transmission consists of one GOP after the other. Typically I frames will have more bits than P and B frames, and B frames will have the least number of bits. The size of I frames can be approximated by a Normal distribution, whereas the size of B and P frames can be approximated either by a gamma or by a lognormal distribution. Figure 3 illustrates the bit rate profile of a typical videostream. The high peaks correspond to scene changes, whereas the low peaks correspond to the activity within a scene.

Video traffic exhibits long-range dependencies [2,8]. The repetitive pattern of GOPs introduces strong periodic components in the autocorrelation function (ACF). Video streams are usually modeled either by a fractal Brownian motion process or by a fractal ARIMA (0,d,0) process, which are LRD processes. However, some researchers advocate that, for finite buffer, long-term correlation have minor impact on queuing performance, and, therefore, Markovian models should be used, since only short term correlations impact the performance. The discrete first-order autoregressive model, DAR(1), is a popular Markovian process used for video modeling. Actually, Markovian models give rise to ACF of the form $\rho(k) \sim e^{-\beta k} (\beta > 0)$, whereas an LRD process exhibits ACF of the form $\rho(k) \sim k^{-\beta} = e^{-\beta \log^k} (\beta > 0)$ [10]. In fact, the performance of fractal models may be overly sensitive to the buffer size, and, consequently, may underestimate the actual performance. On the other hand, Markovian models provide good performance under heavy loads; however, they perform poorly under light loads [10].

3.4. Elastic Sources

The amount of data an application can pump into the network is often regulated by the network protocols and their congestion control mechanisms, which probe the available bandwidth to determine the amount of data that can be transmitted. Traffic sources whose transmission rate depend on network congestion status are called *elastic sources*. Examples of elastic sources are the available bit rate service (ABR) in ATM networks and the Transmission Control Protocol (TCP), largely deployed in the Internet.

TCP congestion control mechanism is a window-based one. Segments, or "packets" in TCP language, are transmitted and acknowledgments from the receiver are expected. Each segment has a sequence number, set at the sending end. Acknowledgments specify the sequence number of the acknowledged segment. Acknowledgments are cumulative; an acknowledgment notifies the transmitter that all the segments with a lower sequence number were properly received. The time from sending a packet to receiving its acknowledgment is called round–trip time (RTT). TCP controls a connection rate by limiting the number of transmitted-but-yet-to-be-acknowledged segments. In the beginning, the window size is set to one. Every time an acknowledgment is received, that is, at every RTT, the window size is doubled, and the window grows up to a threshold. After this threshold, the window is incremented by one segment.

Whenever an acknowledgment fails to arrive after a predefined interval, a timeout event occurs and the threshold is set to one-half the current congestion window and the congestion window is set to one. If the transmitter receives three consecutive acknowledgments for the same segment, it is assumed that the next segment was lost and the window is set to one-half its current value.

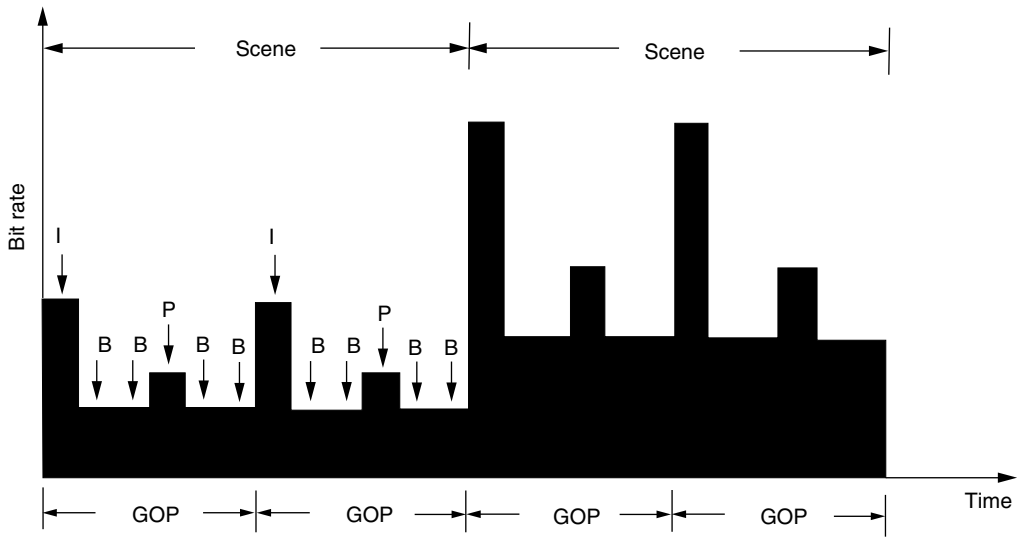


Figure 3. Bit rate of a video transmission.

The evolution of the window size between loss events can be analyzed in order to determine a TCP connection throughput, that is, the amount of data a TCP connection pumps into the network per unit of time by observing the window size evolution between loss events. The distribution of interloss periods as well as the distribution of the type of the loss event should be factored into this computation.

4. EFFECTIVE BANDWIDTHS AND ENVELOPE PROCESSES

Most communications services are subject to performance constraints designed to guarantee a minimal quality of service (QoS). Consider a general traffic stream offered to a deterministic server, and assume that some prescribed parameterized performance constraints are required to hold. The effective bandwidth of the traffic stream corresponds to the minimal deterministic service rate, required to meet these constraints. Queuing-oriented performance constraints include bounds on such statistics as queuing delay quantiles or averages, server utilization, and overflow probabilities. The effective-bandwidth concept serves as a compromise between two alternative bandwidth allocation schemes, representing a pessimistic and an optimistic outlook. The strict one allocates bandwidth based on the stream peak rate, seeking to eliminate losses, whereas the lenient one allocates bandwidth based on the stream average rate, merely seeking to guarantee stability.

Let us formally define effective bandwidth, Let $X[0,t]$ be the workload that arrived during time interval $[0,t]$ for a traffic stream. Effective bandwidth of the traffic stream is defined as $\alpha(s) = \lim_{t\to\infty} \frac{1}{st} \log E(e^{sX[0,t]})$, which is a function of $s > 0$, the so-called space parameter. In the effective bandwidth theory, s is the asymptotic exponential decay rate of queue size distribution tail probability with respect to queue size; that is, when the service rate is $\alpha(s)$, the queue size distribution tail probability with respect to queue size is $P(Q > B) \approx \exp(-sB)$, where Q denotes the queue size. This is why s is called the *space parameter*.

The notion of effective bandwidth provides a useful tool for studying resource requirements of telecommunications services and the impact of different management schemes on network performance. Estimates of effective bandwidths are called *empirical effective bandwidths*, while the analytic form are called *analytic effective bandwidths*.

4.1. Envelope Processes

An envelope process is a function which provides a bound for the amount of work generated in a traffic stream during a certain time interval. If $A(t_2 - t_1)$ is the amount of bits generated during the interval $t_2 - t_1$, then $\hat{A}(t)$ is an envelope process for $A(t)$ if and only if $\hat{A}(t_2 - t_1) > A(t_2 - t_1)$, for any $t_2 > t_1$. For any traffic stream, $A(t)$, there is a whole family of possible envelope processes; however, the lowest bound is the one of interest. Envelope processes are useful tools since, in general, they require a small number of parameters — that is, they are a *parsimonious* way of representing a stochastic process. However, dimensioning based on envelope processes may overestimate the required resources. Moreover, envelope processes are not appropriate for the study of phenomena at the cell timescale, such as cell discarding.

Network services can be either deterministic or statistical. Accordingly, deterministic and stochastic envelope processes are defined. A deterministic envelope process is a strict upper bound on the amount of work arriving during an interval for a traffic stream. A commonly used envelope process is $\int_0^t A(t) < \rho t + \sigma$, where ρ is the source mean arrival rate and σ is the maximum amount of work allowed in a burst [3]. $\rho t + \sigma$ is a model for the output of a leaky bucket regulator where ρ is the leaky rate and σ the bucket size.

In a stochastic envelope process, the amount of work generated in a certain interval may surpass a deterministic bound with a certain probability value. An accurate stochastic envelope process for a fractal Brownian motion process is $\rho t + k\sigma t^H$, where ρ is the mean arrival rate, σ is the standard deviation, and H is the Hurst parameter [6]. Note that the amount of work is not a linear function of time, it has a t^H which takes into account long periods of arrivals.

5. CONCLUSIONS

The aim of traffic modeling is to provide network designers with simple means to predict the network load, and consequently, the network performance. Since the early days of telephony networking, engineers have been engaged in understanding the nature of network traffic, and its impact on quality of service provisioning. Traffic models mimic the traffic patterns observed in real networks. The suitability of a traffic model is related to the degree of accuracy of the conclusions that can be drawn from studies using such a model. Therefore, there is no unique model for a certain type of traffic, but models with different degrees of accuracy.

With the advent of integrated networks, Poisson models for traffic streams were replaced by more sophisticated short range dependent models which considered the correlation pattern besides the mean arrival rate. By 1993, the seminal work of Leland et al. [12] demonstrated the fractal nature of LAN traffic. Several other works followed showing that other types of traffic such as video traffic were also fractal. Recent studies have shown that Internet traffic is not precisely fractal at small timescales, but can be represented well as fractal at larger timescales. The understanding of the impact of multifractality on network performance is still an open problem.

Traffic patterns are influenced by several factors such as the nature of file size, human think time, protocol fragmentation, and congestion control mechanisms. New challenging problems in traffic modeling will certainly exist when multimedia applications become a significant part of the whole network traffic.

BIOGRAPHIES

Mihail (Mike) Devetsikiotis was born in Thessaloniki, Greece. He received the Dipl. Ing. degree in Electrical Engineering from the Aristotle University of Thessaloniki, Greece, in 1988, and the M.Sc. and Ph.D. degrees in

Electrical Engineering from North Carolina State University, Raleigh, in 1990 and 1993, respectively. As a student, he received scholarships from the National Scholarship Foundation of Greece, the National Technical Chamber of Greece, and the Phi Kappa Phi Academic Achievement Award for a Doctoral Candidate at North Carolina State University. He is a member of the IEEE, INFORMS, and the honor societies of Eta Kappa Nu, Sigma Xi, and Phi Kappa Phi. In 1993 he joined the Broadband Networks Laboratory at Carleton University, as a Post-Doctoral Fellow and Research Associate. He later became an Adjunct Professor in the Department of Systems and Computer Engineering at Carleton University in 1995, an Assistant Professor in 1996, and an Associate Professor in 1999. Dr. Devetsikiotis joined the Department of Electrical and Computer Engineering at North Carolina State University as an Associate Professor, in 2000. He has served as an officer of the IEEE Communications Society Technical Committee on Communication Systems Integration and Modeling, and as Associate Editor of the journal *ACM Transactions on Modeling and Computer Simulation*.

Nelson Fonseca received his Electrical Engineer (1984) and M.Sc. in Computer Science (1987) degrees from The Pontificial Catholic University at Rio de Janeiro, Brazil, and the M.Sc. (1993) and Ph.D. (1994) degrees in Computer Engineering from The University of Southern California in Los Angeles. Since 1995 he has been affiliated to the Institute of Computing of the State University of Campinas, Brazil, where is currently an Associate Professor.

He is the recipient of Elsevier Editor of the Year 2000, of the 1994 USC International Book award, and of the Brazilian Computing Society First Thesis and Dissertations award. Mr. Fonseca is listed in Marqui's *Who's Who in the World* and *Who's Who in Science and Engineering*.

He served as Editor-in-Chief for the *IEEE Global Communications Newsletter* (1999–2002). He is an Editor for *Computer Networks*, an Editor for the *IEEE Transactions on Multimedia*, an Associate Technical Editor for the *IEEE Communications Magazine*, and an Editor for the *Brazilian Journal on Telecommunications*.

Dr. Fonseca was the chairman of the 7th IEEE Workshop on Computer-Aided Modeling, Analysis and Design of Communications Networks and Links (CAMAD'98), Vice-Chairman of IEEE GLOBECOM99 Symposium on Multimedia Services and Technology Issues, Vice-Chairman of CAMAD'2000, and Vice-Chairman of the International Teletraffic Congress'17, 2001.

BIBLIOGRAPHY

1. A. Adas, Traffic models in broadband networks, *IEEE Commun. Mag.* (July 1997).
2. J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger, Variable-bit-rate video traffic and long range dependence, *IEEE Trans. Commun.* **43**(2–4): 1566–1579 (1995).
3. R. L. Cruz, A calculus for network delay, part I: Network elements in isolation, *IEEE Trans. Inform. Theory* **37**: 114–131 (Jan. 1991).
4. A. Erramilli, O. Narayan, A. Neidhardt, and I. Sanjee, Performance impacts of multi-scaling in wide area TCP/IP traffic, *Proc. INFOCOM'00*, 2000.
5. A. Feldmann, A. C. Gilbert, W. Willinger, and T. G. Kurtz, Looking behind and beyond self-similarity: On scaling phenomena in measured WAN traffic, *Proc. 35th Annual Allerton Conf. Communications, Control and Computing*, 1997, pp. 269–280.
6. N. L. S. Fonseca, G. S. Mayor, and C. A. V. Neto, On the equivalent bandwidth of self similar sources, *ACM Trans. Model. Comput. Simul.* **10**(3): 104–124 (2000).
7. V. Frost and B. Melamed, Traffic modeling for telecommunications networks, *IEEE Commun. Mag.* (March 1994).
8. C. Huang, M. Devetsikiotis, I. Lambadaris, and A. Kaye, Modeling and simulation of self-similar variable bit rate compressed video: A unified approach, *Proc. SIGCOMM'95 Conf.*, 1995, pp. 114–125.
9. D. Jagerman, B. Melamed, and W. Willinger, Stochastic modeling of traffic processes, in J. Dshalalow, ed., *Frontiers in Queuing: Models, Methods and Problems*, CRC Press, Boca Raton, FL, 1996.
10. M. M. Krunz and A. M. Makowski, Modeling video traffic using M/G/∞ input process: A comparison between Markovian and LRD models, *IEEE J. Select. Areas Commun.* **16**(5): 733–745 (June 1998).
11. N. Laskin, I. Lambadaris, F. Harmantzis, and M. Devetsikiotis, Fractional Lévy motion and its application to traffic modeling, *Comput. Networks,* (*Special Issue on Long-Range Dependent Traffic Engineering*) **40**(3): (Oct. 2002).
12. W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, On the self-similar nature of Ethernet traffic (extended version), *IEEE/ACM Trans. Network.* **2**(1): 1–15 (1994).
13. B. Liu et al., A study of networks simulation efficiency: Fluid simulation vs. packet-level simulation, *Proc. IEEE Infocom*, Alaska, April 2001.
14. B. B. Mandelbrot and J. W. Van Ness, Fractal Brownian motions, fractional noises and applications, *SIAM Rev.* **10**: 422–437 (1968).
15. I. Norros, A storage model with self-similar input, *Queuing Syst.* **16**: 387–396 (1994).
16. K. Park and W. Willinger, eds., *Self Similar Network Traffic and Performance Evaluation*, Wiley-Interscience, 2000.
17. V. Paxson and S. Floyd, Wide area traffic: The failure of Poisson modeling, *IEEE/ACM Trans. Network.* **3**(3): 226–244 (1995).
18. J. A. Silvester, N. L. S. Fonseca, and S. S. Wang, D-Bmap models for the performance analysis of ATM networks, in D. Kouvatsos, ed., *Performance Modeling of ATM Networks*, Chapman & Hall, 1995, pp. 325–346.
19. M. S. Taqqu, V. Teverovsky, and W. Willinger, Is network traffic self-similar or multifractal? *Fractals* **5**: 63–73 (1997).
20. T. Taralp, M. Devetsikiotis, and I. Lambadaris, In search of better statistics for traffic characterization, *J. Braz. Comput. Soc.,* (*Special Issue on Traffic Modeling and Control of Wired and Wireless Networks*) **5**(3): 5–13 (April 1999).
21. S. Tartarelli et al., Empirical effective bandwidths, *Proc. IEEE GLOBECOM 2000*, Vol. 1, 2000, pp. 672–678.
22. B. Tsybakov and N. D. Georganas, Self-similar processes in communication networks, *IEEE Trans. Inform. Theory* **44**(5): (Sept. 1998).

23. W. Willinger and V. Paxson, Where mathematics meets the Internet, *Notices Am. Mathe. Soc.* **45**(8): 961–970 (Sept. 1998).

24. W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, Self-similarity through high-variability: Statistical analysis of Ethernet LAN traffic at the source level, *IEEE/ACM Trans. Network.* **5**(1): 71–86 (Feb. 1997).

NEURAL NETWORKS AND APPLICATIONS TO COMMUNICATIONS

ELIAS S. MANOLAKOS
Northeastern University
Boston, Massachusetts

1. INTRODUCTION

Conventional (serial) computers have a processing unit that executes instructions one after the other, *sequentially*, at a rate that could be as high as 1 billion instructions per second. The way the human brain processes information seems, however, to be quite different. The brain has about 100 billion processing units, called *neurons*, that are highly interconnected. A neuron typically is connected to more than 1,000 other neurons, giving rise to more than 100 trillion connections. Even if only 0.1% of them are active at any given time, and each connection functions as a very slow processor performing one computation every 5 milliseconds, the brain can deliver *in parallel* 100 trillion operations per second (100 Teraops)! Exploiting vast amounts of low-level parallelism may explain why the human brain can perform high-level perceptual tasks, such as visual pattern recognition, speech understanding and so on, very efficiently even though its pattern-searching capabilities are very modest compared with today's fast computers.

Should we try to build computers that work like the human brain? Should we try to imitate closely what we know from biology? These questions have been long debated in the scientific and engineering community. At the one end, *computational neuroscientists* tell us that accurate modeling of the neuronal interactions will illuminate how the brain operates. At the other end, engineers, who are mostly interested in building intelligent machines, tell us that it may be sufficient to draw from the principles that led to successful designs in biology without trying to approximate the biological systems very closely. So starting from the same inquiry, the highly interdisciplinary field of *neural networks* (NN), or *neurocomputing*, has emerged and moved outward in several different directions, and it continues to evolve.

Artificial neural networks (ANNs) are highly interconnected distributed information processing architectures that are built by connecting many simple processing unit models. They come in many different flavors, but all are characterized by their ability to *learn*, that is, to adapt their behavior and structure to capture and form a representation of the process that generates information in the environment where they operate. It is typical to select first an appropriate NN model; then adequately *train* it, using data representative of the underlying environment; and finally present to it novel (unseen) data and let the network extract useful information, a process called *generalization*, or *recall*. Neural networks have become a mainstream information technology and are no longer considered exotic techniques. They are commonly used in data analysis (time series, forecasting, compression, etc.), in studying systems behavior (system identification, adaptive control, chaotic behavior etc.), and in modeling stochastic processes (for uncertainty characterization, pattern classification etc). Neural network solutions have been tried on all types of scientific and engineering problems, ranging from signal and image processing, communications, and intelligent multimedia systems, to data mining, credit card fraud detection, economic forecasting, modeling of ecological systems and detecting patterns in gene expression microarrays, and so on.

This article is organized as follows: In Section 2, we introduce basic knowledge NNs such as the structure of commonly used processing elements, a review of popular network architectures, and associated learning rules. In Section 3, we discuss why different NN architectures have found so many applications in communications by selecting some well-known subareas and discussing representative cases of their use.

2. BASIC NEURAL NETWORKS THEORY

2.1. The Neuron Node

Let us start by considering a simplified view of a bipolar biological neuron. At the "center" of the neuron there is the cell body, or *soma*, that contains the cell *nucleus*. One or more *dendrites* are connected to the nucleus. They are so called because they structurally resemble the branches of a tree. Dendrites form the receiving end of the nerve cell, and their role is to sense activity at their neighborhood and generate a proportional amount of electrical impulses that are sent toward the cell body. A long signal transmission line, called an *axon*, is emanating at the cell body and carries the accumulated activity (action potentials) away from the soma toward other neurons. At the other end of the axon there are *synaptic terminals*. There, the transmitted signal is translated to signals sensed by the dendrites of neighboring neurons through an electrochemical interface process called a *synapse*. A typical cortical neuron may have a few thousand synapses for receiving input, and its axon may have a thousand or so synaptic terminals for propagating activity to other neurons' dendrites.

The communication operations of the simplified neuron model can be summarized as follows: Signals (impulse trains) arrive at the various input synapses. Some signals are stronger than others and some synapses are less receptive than others, due to causes such as fatigue, exposure to chemical agents, and so on. So, as induced signals travel toward the cell body, they effectively are "weighted" (i.e., they are not all deemed equally important). The induced activity is integrated over time and if it exceeds a threshold, the neuron "fires" (produces an output). The generated output signal is transmitted along the axon and may induce signals at the dendrites of other neurons in the proximity.

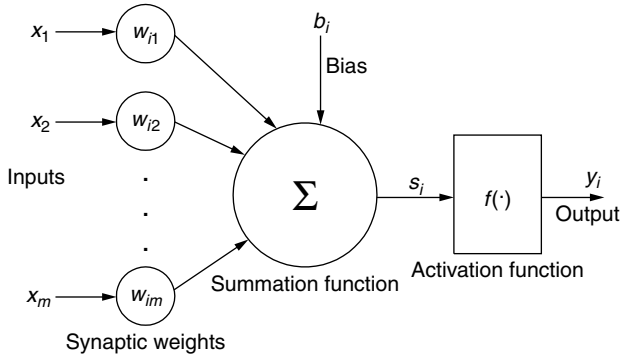


Figure 1. The basic neural network node model.

A neuron node, or *processing element* (PE), of an ANN is a simplistic model that mimics the sequence of the actions described above (see also Fig. 1). Each input $x_k, k = 1, 2, \ldots, m$, to PE_i is first multiplied by the corresponding *synaptic weight*, w_{ik}, that models the strength of the synapse between neurons k an i. Then all weighted inputs are summed. A node specific *bias*, or forcing term, b_i, is also added to the partial sum, providing an additional mechanism to influence node i directly and not through the outputs of other neurons. The *induced local field* s_i becomes the input of an *activation function* $f(\cdot)$ that determines the output y_i of the node. The whole processing can be described by the following two simple equations:

$$s_i = \sum_{k=1}^{m} w_{ik} x_k + b_i \tag{1}$$

$$y_i = f(s_i) \tag{2}$$

The activation, or *transfer function* $f(\cdot)$ usually is nonlinear and limits the range of the values of the neuronal output y_i. Among the most widely used functional forms are those shown in Fig. 2. The hard delimiter (left-most panel) produces a binary, 0 or 1, output when the induced local field has a negative value or positive value, respectively. The transition from level 0 to level 1 is gradual when using a linear ramp-like function (middle panel). Perhaps the most commonly used activation function is the *sigmoidal*

$$f(x) = \frac{1}{1 + e^{-ax}} \tag{3}$$

(see right-most panel). As a increases, the sigmoidal nonlinearity behaves like a hard delimiter. When a -1, $+1$, *antipodal* hard delimiter is used, the neuron node is also usually referred to as a *perceptron*.

2.2. Neural Network Architectures

As with biological NNs, individual neurons are not very useful when working in isolation. Neuron units are usually organized in layers. In a *feedforward* NN, the outputs of the nodes in one layer become inputs to the nodes of the next layer. This is shown in Fig. 3, where circles depict neuron units and arcs represent weighted connections among them. The network is *fully connected* because the output of a node feeds into every node of the next layer

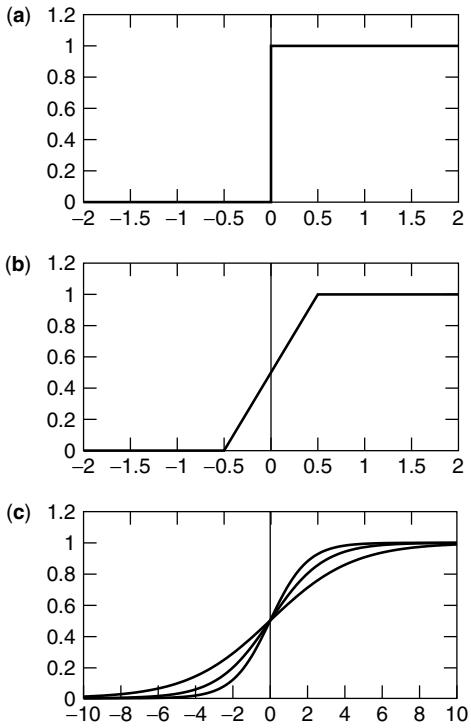


Figure 2. Neural network node activation functions. From left to right: (**a**) hard delimiter; (**b**) piecewise linear; (**c**) sigmoidal; it approaches the hard delimiter as a increases.

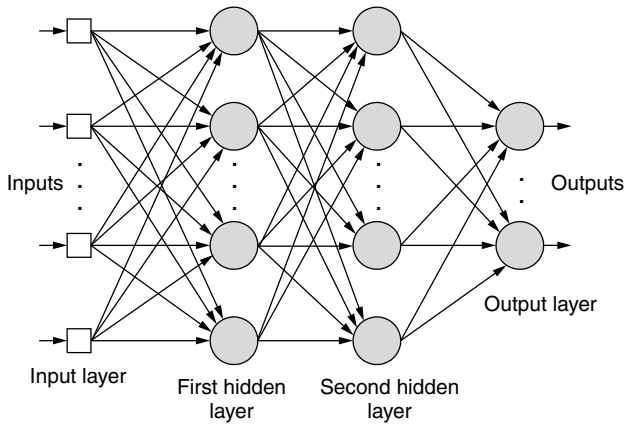


Figure 3. The multilayer feedforward neural network architecture.

via a weighted connection. It has one input layer, two *hidden* layers, and one output layer. The input layer is where the network receives stimuli from the environment, and the output layer is where it returns its response. The middle layers are called "hidden" because their nodes are not directly accessible from outside the network.

If the output of a neuron may become an input to itself, or to other neurons of the same or previous layers, the resulting architecture is called a *feedback*, or *recurrent*, NN. A single layer, fully recurrent NN with four nodes is shown in Fig. 4, where a small box denotes a synaptic weight. Each PE receives a weighted input from the output of every node, including itself, and from an external bias.

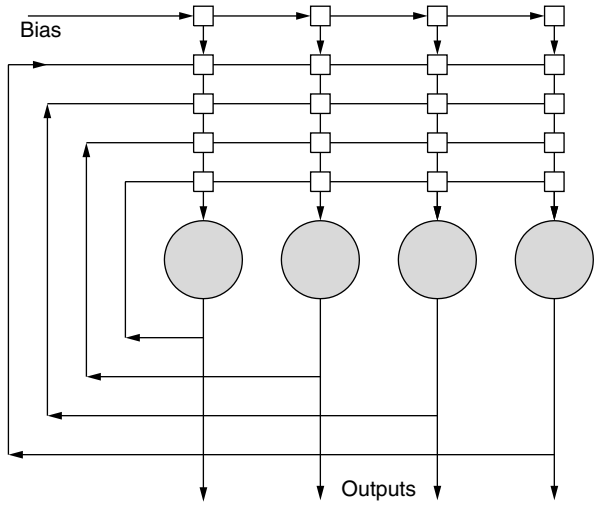


Figure 4. A single layer fully recurrent neural network architecture.

All weighted inputs are accumulated to produce the neuron activation that is passed through a nonlinear thresholding transfer function to produce the new output, as shown in Fig. 1.

Hopfield neural networks (HNNs) [1,2], are single-layer recurrent networks that can collectively provide good solutions to difficult optimization problems. A connection between two neuron processors (analog amplifiers) in established through a conductance weight T_{ij}, which transforms the voltage outputs of neuron unit j to a current input for neuron unit i. Externally supplied bias currents I_i are also feeding into every neuron processor.

It has been shown [3] that in the case of symmetric weight connections ($T_{ij} = T_{ji}$), the equations of motion for the activation of the HNN neurons always lead to convergence to a stable state, in which the output voltages of all the neurons remain constant. In addition, when the diagonal weights (T_{ii}) are zero and the width of the neuron amplifier gain curve is narrow, (i.e., the nonlinear activation function $f(\cdot)$ approaches the antipodal thresholding function), the stable states of a network with N neuron units are the *local* minima of the quadratic *energy function*

$$E = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} T_{ij}V_iV_j - \sum_{i=1}^{N} V_iI_i \tag{4}$$

If the small boxes in Fig. 4 represent conductance weights that meet the conditions discussed, the recurrent NN shown in the figure becomes an HNN with $N = 4$ neuron processors.

2.3. Learning and Generalization

2.3.1. Supervised Learning. Neural network learning amounts to adjusting the synaptic weights periodically and until an appropriate cost function is sufficiently minimized. There are two categories of learning methods. In *supervised* learning, also called "training with a teacher," it is assumed that a desired response $\mathbf{d}$ is provided for every input vector $\mathbf{x}$ in a training set $\mathcal{X}$. For every pair $(\mathbf{x}, \mathbf{d}) \in \mathcal{X}$ the network output $\mathbf{y}$ is compared to $\mathbf{d}$ and the difference (error) vector $\mathbf{e} = \mathbf{d} - \mathbf{y}$ is used to determine how the network weights will get updated, as suggested by the block diagram in Fig. 5 (a).

Supervised learning can be used to train a neural network to act as a *pattern classifier*. Let us assume for simplicity that the network input patterns (vectors) belong to one of two possible classes, C_1 or C_2. Then it suffices to use only one neuron in the output layer; its desired output can be $+1$ if it is known that the pattern belongs to class C_1, or -1 if the pattern belongs to class C_2. By adding more neurons in the output layer, neural classifiers that can discriminate among more than two categories can be built.

Supervised learning can also be used in the more general *pattern association* context. If the input and desirable output patterns are identical, the network is trained to act as an *associative memory*. After adequate training, when the network is presented with an unseen input pattern, it is expected to *recall* the learned pattern that most closely resembles the input. In general, input vectors can be n-dimensional and desirable output vectors m-dimensional, with $m \neq n$.

An example of a weight updating rule used in supervised learning is the so-called *delta rule*, or *Windrow-Hoff* rule. It states that the amount of weight adjustment of synapse k to neuron i, Δw_{ik}, should be proportional to the product of the observed error e_i and the input signal to

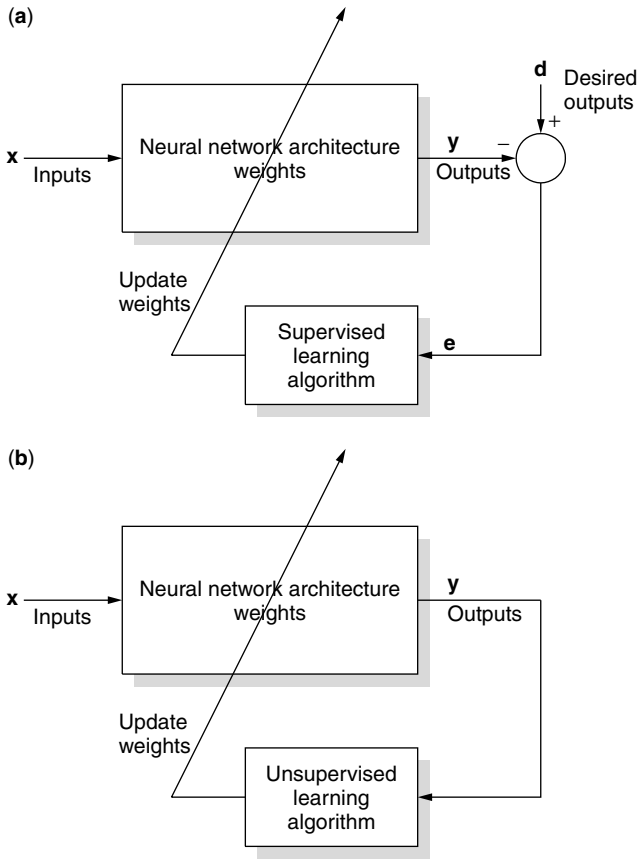


Figure 5. Learning methods: (**a**) supervised training; (**b**) unsupervised training.

neuron i from neuron k, that is

$$\Delta w_{ik} = \alpha e_i x_k \tag{5}$$

Constant α controls the *learning rate* and usually is a small number between 0 and 1. Therefore, after the presentation of input pattern n, the updated weight becomes

$$w_{ik}(n+1) = w_{ik}(n) + \Delta w_{ik}(n) \tag{6}$$

Note that the delta rule uses only information local to neuron i in updating its synaptic weights. However, it does assume that the error term is accessible (i.e., that a desired output d_i is provided for every neuron). It has been shown that if the input vectors in a training set can be separated by a linear hyperplane into two classes, this hyperplane (decision boundary) can be determined by using the weights, upon convergence, of a single layer perceptron trained using correction-based learning.

The error term is directly accessible only for the neurons in the output layer L of a multilayer perceptron (MLP). However, an elegant generalization of the delta rule has been derived that allows the supervised training of MLPs. It is widely known as the *back-propagation algorithm* [4,5] and works as follows. First, the network weights are initialized and an input pattern $\mathbf{x}(n)$ is presented to the input layer. The pattern propagates forward to produce output $\mathbf{y}(n)$ in the output layer L. The network error for pattern n can be defined as

$$E(n) = \frac{1}{2}\sum_{i \in L} e_i^2(n) \tag{7}$$

Adjusting the weight to minimize the total average error $E = 1/N \sum_n E(n)$ leads to a generalized delta rule

$$\Delta w_{ik}(n) = \alpha \delta_i(n) x_k(n) \tag{8}$$

where $\delta_i(n) = \dfrac{\partial E(n)}{\partial s_i}$ is the *local gradient* of the network error for pattern n with respect to neuron i. If neuron i is at the output layer L, this quantity is directly proportional to the error $e_i(n)$. On the other hand, if neuron i is at some level $l < L$, $\delta_i^l(n)$ is proportional to the sum of the $w_{ji}\delta_j^{l+1}(n)$ terms, taken over all neurons j in layer $l+1$ that are connected to neuron i in layer l. So the back-propagation algorithm starts from the output layer L, where vector δ^L is computed directly using the measured error vector $\mathbf{e}$ and is used to update the synaptic weights from layer $L-1$ to layer L. Then these weights and δ^L are used to compute δ^{L-1} and update the synaptic weights from layer $L-2$ to layer $L-1$, and so on. So the error correction propagates backward, toward the input layer, and the weights connecting neurons in two successive layers can be updated in parallel using only localized processing.

Supervised learning methods have been used extensively in communication applications such as channel modeling and identification, adaptive equalization, and so on, as is discussed in more detail in Section 3.

2.3.2. Unsupervised Learning. As with biological neurons, an ANN may have to learn without a "teacher" (i.e., without using a prelabeled training data set). In *unsupervised learning*, as input patterns are presented one after the other, the network should be able to build a meaningful representation of their distribution. Furthermore, it should detect changes in the important characteristics of this distribution and react by adapting its parameters accordingly, as suggested by the block diagram in Fig. 5 (b).

To *self-organize* the network relies on local processing rules that over time help it discover important features in the input patterns and track how they change. A stable representation eventually emerges, for example, a clustering or a topological map arrangement, that models quite well the input patterns' distribution.

Self-organization requires some form of competition, possibly followed by cooperation among neurons in the neighborhood of the winner. In *competitive learning*, neurons start competing when a new pattern is presented to the network. The neuron that is most excited is usually declared the "winner." Neurons in the winner's neighborhood are then cooperating and update their weights according to a local learning rule. For example, in the case of Kohonen's *self-organizing map* (SOM) [5,6], if neuron i wins in the competition for representing pattern $\mathbf{x}(n)$, only neurons j in its topological neighborhood (same region in the lattice of neurons) are allowed to update their weights according to the rule

$$\Delta \mathbf{w}_j(n) = \alpha h_{ji}(n)(\mathbf{x}(n) - \mathbf{w}_j(n)) \tag{9}$$

where function $h_{ji}(n)$ is defined over the winner's neighborhood and its value decreases as the distance of neuron j from the winner neuron i increases. Furthermore, the domain of $h_{ji}(n)$ may shrink over time. In essence, Kohonen's learning rule "moves" toward the input pattern vector, the weight vectors of neurons located in the winner's neighborhood.

A similar situation arises when training *radial basis functions* (RBF) [5,7]. RBFs are three-layer feedforward NNs. The first layer is used to input patterns; every neuron k in the middle (hidden) layer receives the input pattern vector $\mathbf{x}$ and produces an output based on a Caussian bell-shape nonlinear activation function

$$\phi_k(\mathbf{x}, \mathbf{c}_k) = \exp\left(-\frac{1}{2\sigma_k^2}\|\mathbf{x} - \mathbf{c}_k\|^2\right) \tag{10}$$

where $\mathbf{c}_k$ is the *center* associated with the kth neuron. Finally, each neuron j in the output layer is computing a different linear combination of the hidden layer neuron outputs, that is

$$y_j = \sum_k w_{jk}\phi_k(\mathbf{x}, \mathbf{c}_k) \tag{11}$$

A combination of unsupervised and supervised training can be used for RBF networks. The hidden neuron centers may be updated in an unsupervised manner as follows:

$$\Delta \mathbf{c}^* = \beta\,(\mathbf{x}(n) - \mathbf{c}^*) \tag{12}$$

where $\mathbf{c}^*$ is the center of the winning neuron [i.e., the one at smallest distance from the input pattern $\mathbf{x}(n)$]. The weights $\{w_{jk}\}$ may then be updated using error correction supervised learning

$$\Delta w_{jk} = \alpha\,(d_j(n) - y_j(n))\,\phi_k(\mathbf{x}, \mathbf{c}_k) \tag{13}$$

where $d_j(n)$, $y_j(n)$ are the desired and actual response of output layer neuron j to input pattern $\mathbf{x}(n)$ in the training set $\mathcal{X}$.

Unsupervised training is also used extensively in communication applications, such as coding and decoding, vector quantization, neural receiver structures, and blind equalization, to mention a few.

3. NEURAL NETWORK APPLICATIONS TO COMMUNICATIONS

Neural network techniques have traditionally been employed in solving complex problems where a conventional approach has not demonstrated satisfactory performance or has failed to adequately capture the underlying data-generation process. Although they often do improve the performance substantially, it sometimes is hard to understand the decisions they make and explain how these performance gains come about. This has been a point of criticism to the "black box" use of NNs. However, as the field matures and a clear link to Bayesian statistics is established [8], analyzing systematically and explaining how complex networks form successful representations is becoming more and more possible.

In the limited space of this article, we do not intend to provide a comprehensive survey of the numerous applications that NNs have found in communications. We will rather select a few well-known problems in communications and discuss some indicative neurocomputing solutions proposed for them. The selection of the problems to be discussed does not imply that they are more important than others. For a comprehensive review of the literature, the interested reader is referred to the recent article [9] and the edited collections of papers in Refs. 10 and 11.

3.1. Channel Modeling and Identification

Various feedforward neural network architectures, including the multilayer perceptrons and radial basis functions, are *universal approximators* [12,13]. This important property ensures that a neural network of appropriate size can approximate arbitrarily well any continuous nonlinear mapping from an n- to an to m-dimensional space, where $m \neq n$.

Communication channels often exhibit nonlinear, slowly time-varying behavior. So it is natural that NN techniques have been tried for the modeling and identification of communication channels that are used in receiver design, performance evaluation, and so on. Typical examples include satellite channels identification [14] and the modeling of nonlinear time-varying fading channels [15]. In this context, neural networks, with a moderate number of weights (free parameters) that can be updated in parallel, have been shown to provide an attractive alternative to classical nonlinear system identification methods, such as those based on Volterra series [16].

3.2. Channel Equalization

The demand for very high-speed transmission of information over physical communication channels has been constantly increasing over the past 20 years. Communication channels are usually modeled as linear filters having a low-pass frequency response. If the filter characteristics are imperfect, the channel distorts the transmitted signal in both amplitude and delay, causing what is known as *intersymbol interference* (ISI) [17]. As a result of this linear distortion, the transmitted symbols are spread and overlapped over successive time intervals. In addition to noise and linear distortion, the transmitted symbols are subject to other *nonlinear* impairments arising from the modulation/demodulation process, crosstalk interference, the use of amplifiers and converters, and the nature of the channel itself. All the signal processing techniques used at the receiver's end to combat the introduced channel distortion and recover the transmitted symbols are referred to as *adaptive equalization* schemes.

Adaptive equalization is characterized in general by the structure of the equalizer, the adaptation algorithm, and the use or not of training sequences [17]. Linear equalization employs a linear filter, usually with a finite impulse response (FIR) or lattice structure. A recursive least squares (RLS) algorithm or a stochastic gradient algorithm, such as the least mean squares (LMS), is used to optimize a performance index. However, when the channel has a deep spectral null in its bandwidth, linear equalization performs poorly because the equalizer places a highgain at the frequency of the null, thus enhancing the additive noise at this frequency band [17]. Decision feedback (DFE), in conjunction with a linear filter equalizer, can be employed to overcome this limitation. Although DFE and other methods, such as the maximum likelihood (ML) sequence detection [18], are nonlinear, the nonlinearity usually lies in the way the transmitted sequence is recovered at the receiver with the channel model being linear. If nonlinear channel distortion is too severe to ignore, the aforementioned algorithms suffer from a severe performance degradation. Among the many techniques that have been proposed to address the nonlinear channel equalization problem are those in Refs. 19–21, which rely on the Volterra series expansion of the nonlinear channel.

The authors in Ref. 22 have used an MLP feedforward NN structure for the equalization of linear and nonlinear channels. The network is trained to approximate the correct mapping from delayed channel outputs to originally transmitted symbols. It is demonstrated that significant performance improvements can be achieved. A functional-link NN-based DFE equalizer that exceeds the bit error rate performance of conventional DFE equalizers was reported in Ref. 23. For two-dimensional signaling, such as quadrature amplitude modulation (QAM) or phase shift keying (PSK) [17], NNs that can process complex numbers are needed. The authors in Ref. 24 have designed equalizers based on complex-valued radial basis functions and have shown that they can approximate well the decisions

of the optimal Bayesian equalizer. Combining in a loop a DFE equalizer and a self-organizing features map is discussed in Ref. 25. An interesting two-stage equalization strategy is proposed where the DFE compensates for dynamic linear distortions and the SOM compensates for nonlinear ones.

Fully recurrent neural networks (RNNs) have been used in Ref. 26 for both trained adaptation and blind equalization of linear and nonlinear communication channels. Since RNNs essentially model nonlinear infinite memory filters, they can accurately realize with a relatively small number of parameters the inverse of finite memory systems, and thus compensate effectively for the channel-introduced interferences. Their performance is shown to exceed that of traditional equalization algorithms and feedforward NN schemes, especially in the presence of spectral nulls and/or severe nonlinearities. Furthermore, due to the small number of neurons involved, the computational cost of their training may, in practice, be much smaller than that of the MLP-based equalizers of similar performance.

Blind equalization is a particularly useful and difficult type of equalization when training sequences are undesirable or not unfeasible, as, for example, in the case of multipoint communication networks and strategic communications. In the absence of a training sequence, the only knowledge about the transmitted signal is the constellation from which the symbols are drawn. A novel RNN training approach was introduced in Ref. 26 for the blind equalization of nonlinear channels using only a partial set of statistics of the transmitted signal. It is shown that simple RNN structures can equalize linear and nonlinear channels better than the traditional constant modulus algorithm.

3.3. CDMA Multiuser Detection

Code division multiple access (CDMA) is a spectrum-efficient method for the simultaneous transmission of digital information sent by multiple users over a shared communication channel. The spectral efficiency, as well as the antijamming and other attractive properties, make CDMA spread-spectrum techniques useful in a number of communication technologies, including mobile telephony and satellite communications. The wide bandwidth that spread-spectrum CDMA techniques employ enables them to exploit powerful low-rate error-correction coding to further enhance performance. The major limitation of the CDMA techniques however, is the so-called *near–far* problem. When the power of the signals transmitted by the users becomes very dissimilar, the conventional matched-filter detector exhibits severe performance degradation, so more complicated detectors have to be employed.

The optimum centralized demodulation of the information sent simultaneously by several users through a shared Gaussian multiple access channel is a very important problem arising in multipoint-to-point digital communication networks such as radio networks, local area networks, uplink satellite channels, uplink cellular communications, and so on. In CDMA, each transmitter modulates a different signature signal waveform, which is known to the receiver. At the receiver, the incoming signal is the sum of the signals transmitted by each individual user. To demodulate the received signal, we need to suppress the inherent channel noise, often modeled as an additive Gaussian process, and the multiple access interference (MAI).

It has been shown by Verdu et al. [27,28] that for the Gaussian channel, *optimal CDMA multiuser detection* (OMD) can be formulated as the solution of a quadratic integer programming problem that involves the sampled outputs of a bank of filters matched to the signature waveforms of the transmitting users as well as the cross-correlations of them. In Ref. 29, it is proved that, in both the *synchronous* and the *asynchronous* transmission cases, OMD is a computationally expensive problem for which polynomial time solutions most likely do not exist (NP-hard). Therefore, research efforts have concentrated on the development of suboptimal receivers that exhibit good near–far resistance properties, have low computational complexity, and achieve bit-error-rate (BER) performance that is comparable to that of the optimal receiver. Among the many suboptimal multiuser detectors proposed in the literature, we mention the *decorrelating detector* [28], which is linear in nature and complexity and achieves near-optimal performance, assuming that the users' signals form a linearly independent set and the spreading codes of all users are known. Another suboptimal detector is the *multistage detector* (MSD) [30], which relies on improving each stage's estimate by subtracting the estimate of the MAI obtained by the previous stage.

Feedforward NN-based multiuser detectors were first proposed in Refs. 31 and 32. While their performance is shown to be very good for a very small number of synchronous or asynchronous users, their hardware complexity (number of neurons and training time) appears to be *exponential* in the number of users, as conjectured in Ref. 31. Furthermore, it is only empirically possible to determine the number of neurons in the hidden layer as the number of users increases.

The well-known ability of HNNs to provide fast suboptimal solutions to hard combinatorial optimization problems has been exploited to implement efficiently the CDMA OMD in Ref. 33. Starting from the observation that the OMD problem's objective function can be put in the quadratic format of equation (4), an HNN multiuser detector is derived that is proven to be a generalization of the multistage detector. The HNN-based detector has been evaluated via extensive simulations and has been found to outperform the CD by orders of magnitude, exceed the performance of the MSD, and approach the performance of the OMD. Furthermore, the HNN-based detector has a hardware complexity (number of neurons) that is *linear* in the number of users K and does not require any training. Since it can be implemented directly using analog VLSI hardware, its computational cost per symbol is practically constant irrespective of the number of users.

However, as the number of simultaneous users increases, all NN-based receivers may become impractical. To address this severe limitation, a hybrid digital signal preprocessing–NN CDMA multiuser detection scheme was proposed in Ref. 34. An investigation on the nature of the local minima of the OMD's objective function led to the formulation of a computationally efficient digital

signal preprocessing stage that recursively reduces the size of the search space over which the original large-size OMD optimization problem has to be solved. After preprocessing, the remaining optimization problem has the same structure as the OMD but is much smaller and can be solved by an HNN implementable directly in hardware. The lesson learned is that combining conventional DSP with NN methods is worth considering because it often leads to optimized and practical solutions.

3.4. Networking Applications

Multimedia teleconferencing, video-on-demand, and distant learning are only a few examples of emerging applications that require high-speed communications. Each such application presents to the underlying network infrastructure different traffic characteristics and quality of service (QoS) requirements. An intelligent network should efficiently broker among applications with time-varying profiles to meet their short-term requirements while also maximizing the delivered long-term throughput.

The *asynchronous transfer mode* (ATM) is a widely accepted backbone networking technology due to its provisions for QoS delivery. ATM has builtin proactive and reactive mechanisms for effectively managing network resources (e.g., buffer space, etc.) according to traffic profiles and QoS requirements. By using them, it is possible to statistically time multiplex among sources with different burstiness and bit rate characteristics (e.g., voice, data, video) as they compete for the available bandwidth. It is not surprising that NNs, with their learning from examples, adaptation, and prediction capabilities, are among the artificial intelligence methods that have been employed extensively in this context.

The proactive mechanism that decides if a new call can be accepted given the current state of the network is termed *call admission control* (CAC). The integration of adaptive CAC and link capacity control for multimedia traffic over ATM is discussed in Refs. 35 and 36. Neural networks are trained to estimate the cell loss rate from link capacity and observed traffic. The link capacity assignment is then optimized according to the estimated cell loss rate. The application of NNs to the selective admission of a set of calls from a number of inhomogeneous call classes with differing rate and traffic variability characteristics is discussed in Ref. 37.

A good example of adaptive network traffic management is the dynamic allocation of video transmission bandwidth by taking into account the rate of scene changes. In Ref. 38 it is shown that it is feasible to predict scene changes online by employing low-complexity time-delay NNs and use these predictions in dynamic bandwidth allocation for the efficient transmission of real-time video traffic over ATM networks. Furthermore, in Ref. 39 a system consisting of pipelined recurrent NN models, where each RNN has a small number of neurons, is shown to be able to accurately predict the future behavior of MPEG video traffic, based on the ability of each RNN module to learn from previous traffic measurements. The system was used to guide control actions in real time and to prevent excess network loading.

An important task of adaptive network management is to ensure that applications honor their commitments and to respond if they violate their "contract" with the network. During call progress, the same parameters used for call admission may be utilized by a reactive network *policing* mechanism to ensure that the user's traffic remains within the prenegotiated values. A network policing architecture consisting of two interconnected NNs is introduced in Ref. 40. The first NN is trained to learn the probability density function (pdf) of the nonviolating traffic profile of the application. The second NN is trained to capture the characteristics of the pdf of the actual traffic as the call progresses. If the two distribution profiles start deviating considerably, an error signal is generated and used to "reshape" the accepted traffic [41]. The results show that this policing method can efficiently detect and properly react to peak and mean traffic violations. In Ref. 42 a closed loop, end-to-end, broadband network traffic control system is presented that employs different NN architectures for traffic characterization, call admission control, traffic policing, and congestion control.

3.5. Other Communications Applications

Other telecommunication areas where NNs have been used include, but are not limited to: cellular and mobile communications [43,44], mobile phone and credit card fraud detection [45,46], intelligent switching and routing [47,48], scheduling problems in radio networks [48,49], and so on. A large collections of application-related papers can be found in books [10,11].

Neural networks have also been extensively utilized in diverse application domains that are related to communications, such as electronic commerce over the Internet [50], modeling agents behavior [51], automatic language identification [52], text-independent speaker verification [53], character recognition and document analysis [54], image vector quantization [55], signal processing [56], and pattern recognition [8].

4. THE FUTURE?

The highly interdisciplinary field of NNs has always been an area where engineers, statisticians, neuroscientists, and cognitive scientists meet and interact. It is this cross-fertilization of ideas that has rejuvenated NN research over the past 10 years. At the one end, the field has established new links with areas of advanced statistics and machine learning, such as independent component analysis [57], support vector machines [58], graphical models [59], and so on. Coupled with developments in computational neuroscience and digital communications, these advances may soon lead to breakthroughs in novel fields such as that of *neurotechnology*, which promises the use of information technology to substitute for lost functionality in the human nervous system [60].

BIOGRAPHY

Elias S. Manolakos is leading the Parallel Processing and Architectures research group of the Communications

and Digital Signal Processing (CDSP) Center, a world-class Center for Research and Graduate Studies of the Electrical and Computer Engineering Department at Northeastern University, Boston, Massachusetts, where he is currently an associate professor. His research interests are in parallel and distributed computing; embedded systems; pattern recognition; neural networks; and applications in signal processing, communication, and biocomputing. He has served on the editorial boards of the *IEEE Transactions of Signal Processing, IEEE Computing in Science and Engineering, Journal of VLSI Signal Processing, IEEE Letters*, among others Manolakos has participated in the organization of several conferences and has chaired the technical program of the IEEE International Workshop on Signal Processing Systems Design and Implementation (SiPS, 1998) and the IEEE International Workshop in Neural Networks for Signal Processing (NNSP, 1995). He has authored or coauthored with his students more than 70 referenced publications and has coedited three books. He is a senior member of the IEEE and an elected advisory board member of the IEEE Signal Processing Society's Technical Committee on Neural Networks for Signal Processing.

BIBLIOGRAPHY

1. D. V. Tank and J. J. Hopfield, Simple "neural" optimization networks: an A/D converter, signal decision circuit, and a linear programming circuit, *IEEE Trans. Circuits and Systems* **33**: 533–541 (May 1990).
2. J. J. Hopfield, Neural networks and physical systems with emerging collective computational abilities, *Proc. Natl. Acad. Sci. USA* **79**: 2554–2558 (1982).
3. J. J. Hopfield, Neurons with graded response have collective computational properties like those of two-state neurons, *Proc. Natl. Acad. Sci. USA* **81**: 3088–3092 (1984).
4. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Learning internal representation by error propagation. In D. E. Rumelhart and J. L. McClelland, eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1, MIT Press, Cambridge, MA, 1986, pp. 318–362.
5. S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed., Prentice Hall, Upper Saddle River, NJ, 1999.
6. T. Kohonen, Self-organized formation of topological feature maps, *Biological Cybernetics* **43**: 59–69 (1982).
7. F. Girosi, M. Jones, and T. Poggio, Regularization theory and neural networks architectures, *Neural Computation* **7**(2): 219–269 (1995).
8. C. M. Bishop, *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford, UK, 1995.
9. M. Ibnkahla, Applications of neural networks to digital communications — a survey, *Signal Processing* **80**: 1185–1215 (2000).
10. N. Ansari and B. Yuhas, *Neural Networks in Telecommunications*, Kluwer Academic Publishers, Dordrecht, Netherlands, 1994.
11. J. Alspector, R. Goodman, and T. X. Brown, eds., *Applications of Neural Networks to Telecommunications*, Lawrence Erlbaum, Hillsdale, NJ, 1993.
12. G. Cybenko, Approximation by superpositions of a sigmoidal function, *Mathematics of Control, Signals, and Systems* **2**(4): 303–314 (1989).
13. K. Hornik, M. Stinchcombe, and H. White, Multilayer feedforward networks are universal approximators, *Neural Networks* **2**: 359–366 (1989).
14. M. Ibnkahla, N. J. Bershad, J. Sombrin, and F. Castanié, Neural network modeling and identification of non-linear channels with memory: Algorithms, applications and analytic models, *IEEE Trans. Signal Proc.* **46**: 1208–1220 (May 1998).
15. M. Ibnkahla, J. Sombrin, F. Castanié, and N. J. Bershad, Neural network modeling non-linear memoryless communication channels, *IEEE Trans. Commun.* **45**: 1208–1220 (July 1997).
16. M. Schetzen, *The Volterra and Wiener Theories of Nonlinear Systems*, Wiley, New York, 1980.
17. J. Proakis, *Digital Communications*, Prentice Hall Inc., Cliffside Park, NJ, 1988.
18. G. D. Forney, Jr., Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference, *Proc. IEEE, Trans. Infor. Theory* **IT-18**: 378–383 (May 1972).
19. S. Benedetto and E. Biglieri, Nonlinear equalization of digital satellite channels, *IEEE J. Select Areas Commun.* **SAC-1**: 57–62 (Jan. 1983).
20. E. Biglieri, A. Gersho, R. D. Gitlin, and T. L. Lim, Adaptive cancellation of nonlinear intersymbol interference for voiceband data transmission, *IEEE J. Select Areas Commun.* **SAC-2**: 765–777 (Sept. 1984).
21. D. D. Falconer, Adaptive Equalization of Channel Nonlinearities in QAM Data Transmission Systems, *Bell Syst. Tech. J.* **57**(7): 2589–2611 (1978).
22. G. J. Gibson, S. Siu, and C. F. N. Cowan, Application of multilayer perceptrons as adaptive channel equalizers, in *ICASSP Int. Conf. Accoustics, Speech and Signal Proc.*, Glasgow, Scotland, May 1989, pp. 1183–1186.
23. A. Husain, S. Soraghan, and T. Durrani, A new adaptive functional-link neural network-based dfe for overcoming co-channel interference, *IEEE Trans. Commun.* **45**: 1358–1362 (1997).
24. S. Chen, S. McLaughlin, and B. Mulgrew, Complex-valued radial basis function networks, Part I: Network architecture and learning algorithms, *Signal Proc.* **35**(1): 175–188 (Jan. 1994).
25. T. Kohonen, E. Oja, O. Simula, and A. Visa, Engineering application of the self-organizing map, *IEEE Proc.* 1357–1384 (Oct. 1996).
26. G. Kechriotis, E. Zervas, and E. S. Manolakos, Using recurrent neural networks for adaptive communication channel equalization, *IEEE Trans. Neural Networks* **5**(2): 267–278 (March 1994).
27. S. Verdu, Minimum probability of error for asynchronous Gaussian multiple-access channels, *IEEE Trans. Inform. Theory* **32**: 85–96 (Jan. 1986).
28. R. Lupas and S. Verdu, Near-far resistance of multiuser detectors in asynchronous channels, *IEEE Trans. Commun.* **38**: 496–508 (Apr. 1990).
29. S. Verdu, Computational complexity of optimum multiuser detection, *Algorithmica* **4**: 303–312 (1989).

30. M. K. Varanasi and B. Aazhang, Multistage detection in asynchronous code-division multiple access communications, *IEEE Trans. Commun.* **38**: 509–519 (Apr. 1990).

31. B.-P. Paris, B. Aazhang, and G. Orsak, Neural networks for multi-user detection in CDMA communication, *IEEE Trans. Commun.* **40**: 1212–1222 (July 1992).

32. U. Mitra and H. V. Poor, Adaptive receiver algorithms for near-far CDMA, in *PIMRC 92*, pp. 639–644, Oct. 1992.

33. G. I. Kechriotis and E. S. Manolakos, Hopfield neural network implementation of the optimal CDMA multiuser detector, *IEEE Trans. Neural Networks* **7**(1): 131–141 (Jan. 1996).

34. G. I. Kechriotis and E. S. Manolakos, A hybrid digital signal processing — neural network CDMA multiuser detection scheme, *IEEE Trans. Circuits Systems II* **43**(1,2): 96–104 (Feb. 1996).

35. A. Hiramatsu, ATM communications network control by neural networks, *IEEE Trans. Neural Networks* **1**(1): 122–130 (March 1990).

36. A. Hiramatsu, Integration of ATM call admission control and link capacity control by distributed neural networks, *IEEE J. Select. Areas Commun.* **9**: 1131–1138 (Sept. 1991).

37. R. Morris and B. Samadi, Neural network control of communications systems, *IEEE Trans. Neural Networks* **5**(4): 639–650 (July 1994).

38. S. Chong and J. Ghosh, Predictive dynamic bandwidth allocation for efficient transport of real-time VBR video over ATM, *IEEE J. Select. Areas Commun.* **13**(1): 12–23 (Jan. 1995).

39. P. R. Chong and J. T. Hu, Optimal nonlinear adaptive prediction and modeling of MPEG video in ATM networks using pipelined recurrent neural networks, *IEEE J. Select. Areas Commun.* **15**(6): 1087–1100 (Aug. 1999).

40. A. Tarraf, I. Habib, and T. Saadawi, A novel neural network enforcement mechanism for ATM networks, *IEEE J. Select. Areas Commun.* **12**(6): 1088–1096 (Aug. 1994).

41. I. Habib, A. Tarraf, and T. Saadawi, A neural network controller for congestion control in ATM multiplexers, *Computer Networks ISDN Systems* **29**(3): 325–334 (Feb. 1997).

42. A. Tarraf, I. Habib, and T. Saadawi, Intelligent traffic control for ATM broadband networks, *IEEE Commun. Mag.* **33**(10): 76–82 (Oct. 1995).

43. T. Fritsch, Cellular mobile communication design using self-organizing feature maps, in Ben Yuhas and Nirwan Ansari, ed., *Neural Networks in Telecommunications*, Kluwer, Dordrecht, Netherlands, 1994, pp. 211–232.

44. X. M. Gao, X. Z. Gao, J. M. A. Tanskanen, and S. J. Ovaska, Power prediction in mobile communication systems using an OPtimal neural-network structure, *IEEE Trans. Neural Networks* **8**(6): 1446–1455 (Nov. 1997).

45. Y. Moreau and J. Vandewalle, Detection of mobile phone fraud using supervised neural networks: A first prototype, in *International Conference on Artificial Neural Networks 97*, Springer, 1997, pp. 1065–1070.

46. J. R. Dorronsoro, F. Ginel, C. Sánchez, and C. Santa Cruz, Neural fraud detection in credit card operations. *IEEE Trans. Neural Networks* **8**(4): 827–834 (July 1997).

47. T. X. Brown, Neural networks for switching, *IEEE Commun. Mag.* **27**(11): 72–81 (1989).

48. Y. Takefuji, *Neural Network Parallel Computing*, Kluwer Academic Publishers, Boston, 1992.

49. L. Wei and R. Chang, Broadcast scheduling in packet radio networks by Hopfield neural networks, *Information Processing Letters* **63**(5): 271–276 (Sept. 1997).

50. C. Giraud-Carrier and M. Ward, Learning customer profiles to generate cash over the internet, in *Proceedings of the Third International Workshop on Applications of Neural Networks to Telecommunications* (*IWANNT'97*), Lawrence Erlbaum Associates, Publishers, June 1997, pp. 165–170.

51. A. E. Henninger, A. J. Gonzalez, M. Georgiopoulos, and R. F. DeMara, A connectionist-symbolic approach to modeling agent behavior: Neural networks grouped by contexts, *Lecture Notes Comput. Sci.* **2116**: 198 (2001).

52. R. A. Cole, J. W. T. Inouye, Y. K. Muthusamy, and M. Gopalakrishnan, Language identification with neural networks: a feasibility study, in *IEEE Pacific RIM Conference on Communications, Computers and Signal Processing*, Victoria, Canada, June 1989, Piscataway, NJ, 1989. IEEE, pp. 525–529.

53. A. Paoloni, S. Ragazzini, and G. Ravaioli, Predictive neural networks in text independent speaker verification: an evaluation on the SIVA database, in *Proc. ICSLP '96*, Vol. 4, Philadelphia, PA, Oct. 1996, pp. 2423–2426.

54. P. D. Gader et al., Neural and fuzzy methods in handwriting recognition, *Computer* **30**(2): 79–86 (Feb. 1997).

55. R. Lancini, Image vector quantization by neural networks, in Ben Yuhas and Nirwan Ansari, ed., *Neural Networks in Telecommunications*, Kluwer Academic Publishers, Dordrecht, Netherlands, 1994, pp. 287–303.

56. B. H. Juang, S. Y. Kung, and C. A. Camm, eds., *Neural Networks for Signal Processing: Proceedings of the 1991 IEEE Workshop*, IEEE Press, 1991.

57. A. Hyvärinen and E. Oja, Independent component analysis: algorithms and applications, *Neural Networks* **13**(4–5): 411–430 (2000).

58. T. Evgeniou and M. Pontil, Support vector machines: theory and applications, *Lecture Notes Comput. Sci.* **2049**: 249–259 (2001).

59. Steffen L. Lauritzen, *Graphical Models*, Clarendon Press, Oxford, UK, 1996.

60. R. Eckmiller, Towards learning retina implants for partial compensation of retinal degenerations, in Dan Lundh, Bjorn Olsson, and Ajit Narayanan, eds., *Biocomputing and Emergent Computation*, World Scientific, 1997, pp. 271–281.

NONLINEAR EFFECTS IN OPTICAL FIBERS

ANDREW R. CHRAPLYVY
Bell Laboratories
Lucent Technologies
Holmdel, New Jersey

1. NONLINEAR EFFECTS IN OPTICAL FIBERS

The field of nonlinear optics in silica optical fibers originated in the late 1960s–early 1970s [1]. Initially nonlinear effects in single-mode silica fibers were laboratory curiosities requiring powerful lasers for their observation. The

discovery of erbium-doped fiber amplifiers for the 1.5-μm wavelength region [2,3] fundamentally altered the lightwave communication landscape by ushering in the era of wavelength-division multiplexing (WDM) and elevating optical nonlinearities to a primary systems consideration.

Before the early 1990s long-haul high-speed digital lightwave systems typically transmitted one wavelength channel on each optical fiber. These signals required frequent (every 40–50 km) 3R regeneration (reamplify, retime, reshape) that was accomplished using optoelectronic regenerators. In principle many information channels, each at a separate wavelength, could be transmitted over a single fiber. This is known as *wavelength-division multiplexing* (WDM). However, at each regenerator site the WDM channels must be optically demultiplexed, individually regenerated, and then optically multiplexed onto the next fiber span. For a large number of channels, this is prohibitively expensive. The advent of erbium-doped fiber amplifiers, which provide broadband optical gain in the wavelength region of minimum loss of silica fibers (1.55 μm), eliminated this problem. The broad amplifier gain bandwidth (~35 nm) allows simultaneous amplification of many WDM channels, thereby eliminating the need for demultiplexing at each repeater site. By replacing 3R regenerators with optical amplifiers, the distance between optoelectronic signal regeneration was increased from about 50 km to hundreds or even thousands of kilometers. However these major benefits of amplifiers increase the effects of optical nonlinearities. WDM increases the optical power propagating through fibers, and replacing regenerators with amplifiers increases the distances between signal regeneration. Increased optical powers and longer interaction lengths magnify the effects of nonlinearities. In WDM systems with several tens of channels propagating several hundreds of kilometers between regenerator sites various optical nonlinearities can be easily observed even though the power in the individual wavelength channels is on the order of 1 mW.

A number of nonlinearities in silica fibers can impact amplified lightwave systems [4]. They fall into two general categories. Stimulated scattering such as stimulated Brillouin scattering and stimulated Raman scattering are interactions between optical signals and acoustic or molecular vibrations in the fiber. Although both processes can produce exponential optical gain, they are qualitatively very different and affect lightwave systems in different ways. The second category of nonlinearities arises from modulation of the refractive index of silica by intensity changes in the signal. This gives rise to nonlinearities such as *self-phase modulation*, whereby an optical signal alters its own phase and spectrum; *cross-phase modulation* in WDM systems, where one optical signal affects the phases and spectra of all other optical signals and vice versa; and *four-photon mixing*, whereby WDM signals interact to produce mixing sidebands (as in intermodulation distortion).

1.1. Stimulated Scattering

1.1.1. Stimulated Brillouin Scattering. *Stimulated Brillouin scattering* (SBS) is the interaction between light and acoustic waves. In optical fibers SBS has the lowest threshold power of all the nonlinearities [5]. Light scatters from acoustic phonons and is downshifted in frequency. The magnitude of the downshift depends on the scattering angle (varying from a zero-frequency shift for forward scattering to a maximum frequency shift in the backward direction). In single-mode silica fibers the only frequency-shifted scattered light that continues to be guided is backward-propagating light. The Brillouin shift for backward scattering in silica is about 11 GHz. This backscattered light experiences exponential gain due to the forward-propagating light. System impairment occurs when the backscattered light level becomes comparable to the signal power and begins to deplete the signal. For typical fibers the threshold power for this process is ~10 mW for single fiber spans and correspondingly lower for concatenated amplified spans. However, optical amplifiers usually have optical isolators (otherwise long amplified systems could easily optically oscillate), which prevent backward SBS light from propagating through multiple fiber spans. Consequently the SBS impairment in amplified systems occurs at the same power levels as in regenerated systems. In addition, SBS impairments are not exacerbated in WDM systems, because each signal channel interacts with acoustic phonons having slightly different frequencies. Thus the nonlinearities accumulate individually for each channel. However, some systems will require individual signal powers greater than 10 mW. For example, to overcome fiber attenuation in extremely long single-span systems, higher signal powers, which may exceed the SBS threshold, are required.

Although SBS has the lowest threshold of all the fiber nonlinearities, it is also the easiest nonlinearity to counteract because of the lifetime of the acoustic phonons that give rise SBS. The phonon lifetime in silica fibers is about 15 ns, which corresponds to an optical linewidth of about 20 MHz. Optical sources with linewidths greater than 20 MHz will experience reduced SBS. Typical laser diodes have linewidths less than 10 MHz, but dithering the laser injection current can artificially broaden the effective linewidths because it causes a dithering of the optical frequency of the signal. The SBS gain is then reduced by the ratio of the magnitude of the frequency dither divided by 20 MHz. It is easy to increase the SBS threshold by an order of magnitude simply by dithering the diode laser frequency over a 200-MHz range. Typically this corresponds to a dither current of about 0.2 mA. The dithering technique is now an industrywide standard whenever SBS is a nuisance.

1.1.2. Stimulated Raman Scattering. *Stimulated Raman scattering* (SRS) in optical fibers is the interaction between light and the vibrational modes of silica molecules in the core of the fiber. Although both SRS and SBS are examples of stimulated scattering, there are a number of key differences between the two nonlinearities. Unlike SBS, SRS is an extremely broadband effect. The Raman transitions in silica glass are very broad and overlap into a continuous-gain curve such as that shown in Fig. 1. Note that the peak SRS gain occurs at a frequency about 15 THz lower than the input signal. Unlike SBS, SRS occurs in both forward and backward directions. Isolators at amplifier sites will not diminish forward SRS, and the effect accumulates with the number of amplified fiber spans.

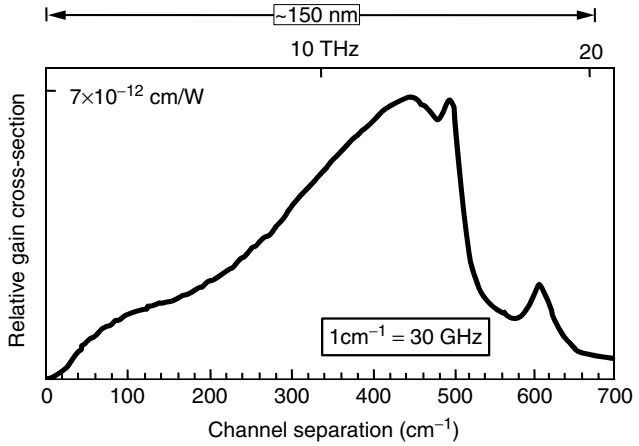


Figure 1. The relative SRS gain coefficient for fused-silica fibers at 1.5 μm.

In single-channel systems some of the spontaneously scattered Raman-shifted light, having amplitude given by the Raman gain curve in Fig. 1, will be guided in the core of the fiber in the forward direction. The copropagating signal light will amplify this scattered light. Because the peak SRS gain is over two orders of magnitude smaller than the peak SBS gain, significantly higher optical powers are required to exceed the SRS threshold for a single channel. The SRS threshold for single fiber spans is over 1 W and manifests itself by exponential amplification of wavelengths near the peak of the Raman gain curve about 15 THz lower in frequency than the signal. In amplified systems one might expect that the threshold is 1 W divided by the number of spans. However, the optical gain bandwidth of an erbium amplifier is roughly 4 times smaller than the bandwidth of the SRS gain profile. Consequently, only Raman light generated within the amplifier gain profile will propagate through the amplified chain of fibers. Since the SRS gain profile is roughly triangular, the peak Raman gain at 30 nm is roughly $\frac{1}{4}$th that of the maximum gain. It follows that in the worst case, the SRS threshold for an amplified chain will be about 4 W divided by the number of amplified spans between regenerators. SRS can be easily suppressed in single-channel amplified systems by periodically inserting optical bandpass filters that pass the signal and reject most of the SRS spectrum.

It is in WDM systems that SRS can be particularly vexing. Because the SRS gain is so broad, WDM channels will be coupled to each other for channels spaced up to 20 THz (150 nm). The short-wavelength channels will act as Raman pumps for long-wavelength channels [6,7]. The long-wavelength channels will be amplified at the expense of the short-wavelength channels, which will be attenuated. Impairments from such interactions will occur at powers much lower than 1 W. For example, for two channels separated by 15 THz (110 nm), unacceptable system degradations will occur at 50 mW in a single fiber span. For multiple channels and multiple spans the threshold powers for degradation will be proportionately smaller. Ultimately, SRS limits the number WDM channels that can be transmitted through single-mode fibers. The number of channels is inversely proportional to the overall transmission distance. Thus far the discussion is based on the assumption of continuous-wave (CW) power at all the signal wavelengths. However, digital systems typically transmit binary information in which a pulse of light represents a logical "one" (called a "mark") and absence of optical power represents a logical "zero" (called a "space"). A space can neither experience Raman gain from shorter wavelengths nor produce Raman gain for longer-wavelength signals. Only marks can be amplified or depleted by SRS, and the amount of gain or depletion will depend on the presence of marks in other channels. Since the occurrence of marks is a random process the amount of amplification or depletion for marks in a particular channel will vary from mark to mark. This is called *pattern-dependent SRS*. Marks with pattern-dependent amplitudes give rise to intersymbol interference (ISI), one of several types of degradations in digital systems. Pattern-dependent SRS is somewhat ameliorated by two effects: modulation statistics and chromatic dispersion. In a binary bit stream the probability of occurrence of marks and spaces is $\frac{1}{2}$. In a WDM system with many channels the occurrence of marks and spaces at a particular instant in time is a random variable. As the number of WDM channels increases, the pattern dependent variation of SRS decreases. This averaging effect is further increased due to chromatic dispersion; specifically, different wavelengths in an optical fiber propagate with different group velocities. Consequently a particular bit in a given channel "samples" multiple time slots (bits) of neighboring channels as it propagates through the fiber. This further diminishes pattern dependent SRS effects. Ultimately this leads to a very efficient method to combat SRS. Filters placed periodically along the fiber (conveniently located within the optical amplifiers themselves) with the inverse filter profile of the SRS gain curve in Fig. 1 can almost completely undo the effects of SRS; specifically, slightly more attenuation is provided to the long-wavelength channels than to the short-wavelength channels. This method of combating SRS is now routinely used in long WDM systems consisting of many channels.

In the early days of WDM systems, optical nonlinearities were viewed as detrimental phenomena that needed to be mitigated. With time clever applications of fiber nonlinearities have led to useful optical devices. An important example is the use of SRS to provide optical gain for the WDM signals. Injecting pump light of the appropriate wavelength(s) into a fiber can turn the fiber into a stimulated Raman amplifier; thus the transmission medium is also an amplification medium. In fact, systems can be designed so that the amount of optical gain produced by SRS exactly compensates the intrinsic attenuation of the transmission fiber. Such a system no longer requires discrete amplifiers every 80–100 km but requires only periodic injection of pump light. The noise figures of Raman-amplified systems are typically several decibels lower than the noise figures of conventional systems amplified by discrete amplifiers. Typically pump light is injected in the backward direction relative to the signals but there are examples of both counter and copropagating Raman pumping.

1.2. Nonlinear Refractive Index

The refractive indices of many optical materials are weakly intensity-dependent ($n = n_0 + n_2 I$). The intensity-dependent refractive index, n_2, of silica has a value of 2.6×10^{-20} m^2/W. Although the n_2 of silica is extremely small (many semiconductor materials have values of n_2 orders of magnitude larger than those of silica), in long amplified systems or in certain WDM systems the effects of the nonlinear refractive index can be quite prominent.

1.2.1. Self-Phase Modulation. *Self-phase modulation* (SPM) describes the effect of a pulse on its own phase [8]. The edge of an optical pulse represents a time-varying intensity. A time-varying intensity in a medium with an intensity-dependent refractive index will produce a time-varying refractive index, which, in turn, produces a time-varying phase that corresponds to a spectral broadening (Fig. 2). Therefore one of the consequences of the nonlinear refractive index of silica is that the spectral width of signal pulses will gradually increase as they propagate in a fiber. For example, a 1-mW pulse would exhibit a twofold broadening after propagating several thousand kilometers in an amplified system. However, a 10-mW pulse will experience the same spectral broadening in 10 times less interaction length.

Spectral broadening can degrade systems in several ways. In a densely spaced WDM system SPM can broaden the signals so that adjacent channels begin to partially overlap spectrally. This leads to optical crosstalk. Spectral broadening can also lead to pulseshape deformation. Because of the nonlinear refractive index, the peak of a pulse accumulates phase more quickly than the wings. This results in stretching of the wavelength on the leading edge of the pulse and compression on the trailing edge. Thus the trailing edge of a pulse acquires a blue shift and the leading edge acquires a red shift (Fig. 3). Recall that most fibers have finite chromatic dispersion. Thus two edges of a pulse will propagate at different speeds. In fibers

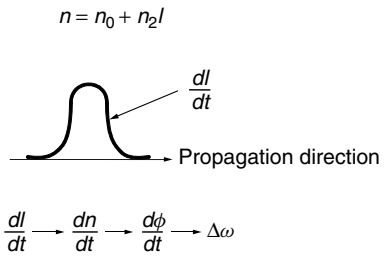


Figure 2. Source of spectral broadening due to nonlinear refractive index.

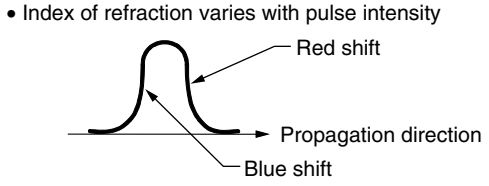


Figure 3. Effect of SPM on leading and trailing edges of a pulse.

with normal chromatic dispersion (red propagates faster than blue) the pulse will begin to temporally broaden. With sufficient temporal broadening pulses will begin to overlap neighboring time slots and interfere with neighboring bits. Bit errors can occur when a mark spreads into a neighboring space. In fibers with anomalous dispersion (blue propagates faster than red) the pulses initially become narrower in time. Ultimately the two edges pass through the center of the pulse, and with further propagation, the pulse will temporally broaden and bit errors will occur when pulses overlap neighboring time slots. System degradations due to the combined effects of SPM and chromatic dispersion are now mitigated by a technique known as *dispersion management*. An optical line system can be designed to consist of two different types of fibers, one having normal dispersion and the other having anomalous dispersion (the transmission fiber itself can be a concatenation of the two types of fiber, or in the more popular design there is one type of transmission fiber and the other type of fiber, called *dispersion compensating fiber*, is located within the optical amplifiers). The fibers are chosen so that the overall accumulated chromatic dispersion for the entire system is nearly zero. In such situations there is no net temporal broadening of the self-phase-modulated pulses, albeit during propagation the pulses "breathe" — broaden as a result of the dispersion of one fiber and then narrow to approximately the original widths due to the opposite-sign dispersion in the second type of fiber.

SPM, like SRS, is a nonlinearity that can be exploited to advantage. Even in the absence of nonlinearity (e.g., at very low powers), pulses propagating in fibers broaden as a result of chromatic dispersion. Each pulse intrinsically contains a spread of wavelengths determined by the pulsewidth and other factors. These wavelength components travel at different speeds, and this leads to pulse broadening. In anomalous dispersion fibers we have seen that the consequence of SPM is pulse narrowing. This tendency to narrow can exactly balance out the broadening due to linear chromatic dispersion and can produce pulses that do not change shape as they propagate. Such pulses are called *solitons* [4]. Soliton technology is now finding its way into commercial transmission systems.

1.2.2. Cross-Phase Modulation. In WDM systems the intensity variations in any signal channel will affect the phases of all the other signals [9]. The origin of this cross-phase modulation (CPM) is the same nonlinear refractive index that gives rise to SPM. If the fiber chromatic dispersion were zero for all channels (all pulses propagate in lock step), the effects of CPM due to each interfering channel would be exactly twice as strong as the SPM effect. However, there are no practical "dispersion-flattened" zero-dispersion fibers. Consequently, the group velocities of various channels in a WDM system are different and pulses in different channels will pass through each other while propagating in the fiber. Under some conditions these pulse collisions virtually eliminate spectral broadening due to CPM. Figure 4 schematically depicts pulses from two different channels passing through each other (the figure is shown in the reference frame of

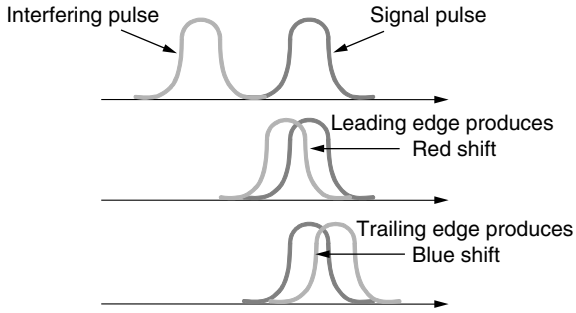


Figure 4. Effects of CPM on colliding pulses.

the signal pulse). Note that during the first half of the collision the interfering pulse produces a red shift in the signal pulse. In the second half of the collision the trailing edge of the interfering pulse produces a blue shift in the signal pulse. The blue shift exactly reverses the effects of the red shift if the intensities and shapes of the pulses have not significantly changed during the collision. Designing transmission systems around this cancellation effect is impractical for two main reasons:

1. A minimum wavelength spacing between neighboring channels is required for the pulse collisions to occur rapidly enough that their intensities do not appreciably change during the collision. This sacrifices precious spectral efficiency (channel bit rate divided by channel spacing) in many cases.
2. There is no way to avoid "partial" collisions. For example, two pulses in neighboring channels exit an optical amplifier partially overlapped. The leading-edge/trailing-edge symmetry is destroyed.

Partial collisions will frequency shift one part of a pulse relative to the remainder. This leads to different group velocities for different parts of a pulse. The occurrence of partial collisions is a random process depending on the presence or absence of marks in various channels. Consequently the arrival time of various marks in a particular signal channel will be random, leading to timing jitter at the receiver, another type of degradation in digital systems.

As with SPM, dispersion management can reduce the effects of CPM. Spectral broadening or frequency shifts do not give rise to dispersion penalties or timing jitter if the overall system dispersion is nearly zero.

1.2.3. Four-Photon Mixing. A third manifestation of the nonlinear refractive index in WDM systems is *four-photon mixing* (FPM). In the case of two signals (Fig. 5a) there exists an intensity modulation at the beat frequency that modulates the refractive index, producing a phase modulation at the difference frequency. This phase modulation creates two sidebands. These sidebands are called *two-tone products* because they were produced by the mixing of two signal waves. For three channels (Fig. 5b), in addition to the two-tone products created by each pair of signals, there are 3 three-tone products generated by all

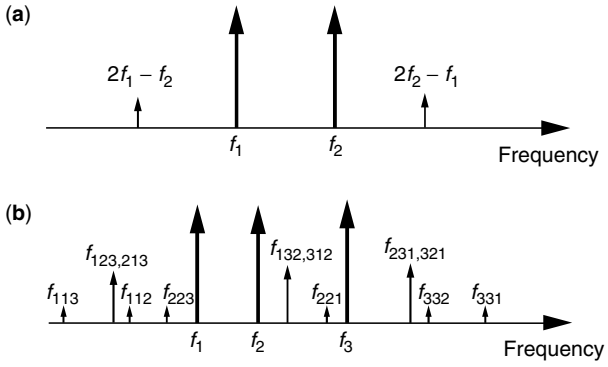


Figure 5. Phase modulation sidebands produced by FPM for (**a**) two signals and (**b**) three signals.

three signals (i.e., the beat frequency of one pair of signals produces sidebands on the third signal). Since there are two different ways of generating each three-tone electric field, the three-tone products are generated with four times the optical power of the two-tone products. For N channels there will be $N^2(N-1)/2$ mixing products generated. For example, in an 8-channel WDM system, 224 mixing products are generated.

Two different impairments are caused by FPM. The obvious and more benign degradation is depletion of signal power in creating the mixing products, especially in WDM systems with many channels. An even more serious degradation occurs if the signal channels are equally spaced. In this case many of the mixing products are produced with optical frequencies the same as the signal frequencies. These mixing products interfere coherently with the signals, either constructively or destructively depending on the (time-dependent) relative phases of the signals. Because it is the electric fields that interfere, small mixing products can produce severe degradations. For example, a mixing product with 1% of the power of one of the signals has an electric field magnitude 10% of the signal electric field and will produce 20% depletion in the signal channel if it destructively interferes with the signal.

To eliminate coherent mixing, a frequency allocation scheme for signals has been devised to ensure that no mixing product will have the same frequency as any signal [4]. This is accomplished by ensuring that the frequency spacing between any two signals is unique. This eliminates interference impairments and leaves only the less severe depletion effects with which to contend. The frequency allocation scheme requires control of signal frequencies to within several gigahertz, easily achievable with present transmitter lasers.

Since FPM is a phase-matched process, it depends strongly on chromatic dispersion [4]. In general, the intensity modulation generated by the beating between two or three signals propagates at a different speed than the signals themselves. If the difference in propagation velocities is large (large chromatic dispersion), the FPM generation efficiency becomes small (poor phase matching) and system degradations are inconsequential. In fibers with zero chromatic dispersion near the signal wavelengths, FPM is a very efficient nonlinear process and dramatic system degradations can occur in relatively short (≅20-km)

lengths of fiber. Dispersion management was initially invented in 1993 to combat the effects of FPM in high-speed systems that were sensitive to chromatic dispersion. The high local dispersion in any of the fiber segments suppresses FPM, but the overall dispersion is nearly zero, to avoid dispersion penalties for high-bit-rate signals. Subsequently dispersion management proved to be a useful technique in counteracting SPM and CPM effects.

FPM can also be exploited for useful purposes, namely, for parametric optical amplification [10]. Signals can be injected into a low-dispersion optical fiber along with a strong pump at a wavelength near the zero-dispersion wavelength of the fiber. FPM between the pump and the signals will produce strong FPM products located in frequency-reversed order (phase conjugation) at wavelengths on the opposite side relative to the pump. The noise figure (NF) of parametric amplifiers can be significantly smaller than the NF of conventional erbium-doped fiber amplifiers, and phase conjugation of amplified signals also reverses the effects of chromatic dispersion. In principle, a system based on parametric amplification would not need dispersion management.

2. CONCLUSION

In conclusion, silica optical fibers exhibit a rich collection of optical nonlinearities that have become important with the advent of practical optical amplifiers and the ultra-long-haul lightwave systems they enabled. Since the early 1990s an arsenal of techniques have been developed to mitigate the effects of nonlinearities. These techniques include dispersion management, frequency or phase modulation of sources, and optical filtering. However, the most tangible impact of optical nonlinearities is to provide gainful employment for systems engineers trying to maximize the ultimate information-carrying capacity of optical fibers by either counteracting or exploiting nonlinear effects in fibers. More advanced treatment of optical nonlinearities in fibers can be found in Chapter 8 of the article by Stolen and Lin [4] and references cited therein.

BIOGRAPHY

Andrew R. Chraplyvy received the B.S. degree in physics in 1972 from Washington University, St. Louis, Missouri, and the M.S. and Ph.D. degrees in physics from Cornell University in 1975 and 1977, respectively. He joined the Physics Department at General Motors Research Labs in 1977 as a Research Scientist. At GM he worked on ultra-high-resolution spectroscopy of gases and impurity modes in solids. Since 1980, he has been with Bell Laboratories, where he currently is Director of Lightwave Systems Research. Dr. Chraplyvy holds over 25 patents in the areas of lightwave systems and fiberoptics. He is the recipient of the 1999 Thomas Alva Edison Patent Award and the 1999 New Jersey Inventor of the Year Award. He is a Bell Labs Fellow, a member of the National Academy of Engineering and a Fellow of the Optical Society of America. His areas of interest are fiberoptics, lightwave communications systems, nonlinear optical interactions in fibers, fiber networks, and high-resolution spectroscopy of gases and solids.

BIBLIOGRAPHY

1. R. H. Stolen, E. P. Ippen, and A. R. Tynes, Raman oscillation in glass optical waveguide, *Appl. Phys. Lett.* **20**: 62–64 (1972).
2. R. J. Mears, L. Reekie, I. M. Jauncey, and D. N. Payne, Low-noise erbium-doped fiber amplifier operating at 1.54 μm, *Electron. Lett.* **23**: 1026–1027 (1987).
3. E. Desurvire, J. R. Simpson, and P. C. Becker, High-gain erbium-doped traveling-wave fiber amplifier, *Opt. Lett.* **12**: 888–890 (1987).
4. I. P. Kaminow and T. L. Koch, eds., *Optical Fiber Telecommunications IIIA*, Academic Press, San Diego, 1997.
5. D. Cotter, Observation of stimulated Brillouin scattering in low-loss silica fiber at 1.3 μm, *Electron. Lett.* **18**: 495–496 (1982).
6. A. R. Chraplyvy and P. S. Henry, Performance degradation due to stimulated Raman scattering in wavelength-division-multiplexed optical-fiber systems, *Electron. Lett.* **19**: 641–642 (1983).
7. A. R. Chraplyvy, Optical power limits in multichannel wavelength-division-multiplexed systems due to stimulated Raman scattering, *Electron. Lett.* **20**: 58–59 (1984).
8. R. H. Stolen and C. Lin, Self-phase modulation in silica optical fibers, *Phys. Rev. A* **17**: 1448–1453 (1978).
9. A. R. Chraplyvy and J. Stone, Measurement of crossphase modulation in coherent wavelength-division multiplexing using injection lasers, *Electron. Lett.* **20**: 996–997 (1984).
10. R. H. Stolen and J. Bjorkholm, Parametric amplification and frequency conversion in optical fibers, *IEEE J. Quantum Electron.* **QE-18**: 1062–1072 (1982).

NONUNIFORMLY SPACED TAPPED-DELAY-LINE EQUALIZERS FOR SPARSE MULTIPATH CHANNELS

FREDERICK K. H. LEE
PETER J. MCLANE
Queen's University
Kingston, Ontario, Canada

1. INTRODUCTION

The high data rate requirement in current and future broadband wireless communication systems has created a new regime of interesting and challenging problems for communication engineers of the new century. Of major concern in the physical layer is the growth of the lengths of the sampled channel impulse responses as a function of the transmission rate when measured in units of symbol intervals. Most commonly used equalization techniques for suppressing intersymbol interference (ISI) distortion caused by the multipath propagation phenomenon, such as the tapped-delay-line (TDL) equalizers, including the linear equalizers (LEs) and the decision feedback equalizers

(DFEs), and the maximum-likelihood sequence estimators (MLSEs), all exhibit structural and computational complexities that depend on the sampled channel lengths. An increase in transmission rate thus inevitably leads to an increase in complexity of these equalization techniques, which is particularly problematic for mobile receivers where resources are scarce because of constraints in cost, size, and battery power. While an impulse response can certainly be truncated to reduce its length, this would be undesirable if its tail portion contains a significant amount of energy. This is precisely the dilemma facing the family of wireless channels called *sparse multipath channels*.

2. SPARSE MULTIPATH CHANNELS

A sparse multipath channel is characterized by an impulse response consisting of only a few dominant multipath terms, but with any two terms separated by a large time delay, resulting in a long delay spread. Examples of sparse multipath channels include terrestrial broadcasting channels, such as those found in high-definition television (HDTV) systems; horizontal or vertical underwater acoustic channels, where reflections off the sea surface and the sea floor constitute the two main causes for the long reverberation of the multipath terms [1, Chap. 8]; as well as cellular land mobile radio channels encountered in hilly terrain environments. Combined with a high data rate, the lengths of the sampled channel impulse responses can range from several tens to hundreds of symbol intervals, depending on the specific application. For instance, in the proposed North American HDTV terrestrial broadcasting mode, 64-QAM (quadrature amplitude modulation) is used and the transmission rate is 5.38 Msps (megasymbols per second). With a typical delay spread of 20 μs, the sampled channel length spans 107.6 symbol intervals, but only a small number of the channel taps exhibit large magnitude due to the sparse nature of the channel [2]. An additional feature of these HDTV channels is the existence of strong precursor taps known as "ghost" signals, which adds to the difficulty of equalization. As another example, Fig. 1 shows the impulse response of a sparse underwater acoustic channel with delay spread that exceeds 80 ms. Hence, even with a modest transmission rate of 1.25 ksps, the sampled channel length is at least a hundred symbol intervals long.

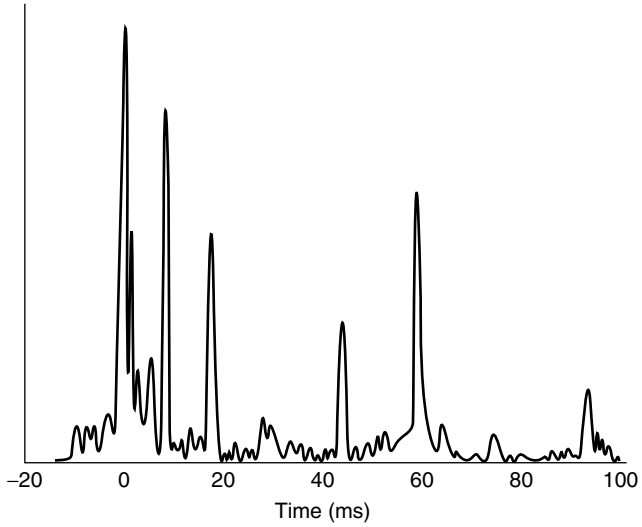


Figure 1. Impulse response of a sparse underwater acoustic channel. (*Source*: Ref. 7, Fig. 1(b).)

The description above clearly explains why existing equalization techniques cannot efficiently mitigate the ISI distortion intrinsic to a sparse multipath channel. Fortunately, as the majority of taps in a sampled sparse multipath channel contain zero or near-zero values, it is possible to develop new equalization methods with complexity associated with the number of large-magnitude taps instead of the entire channel length. One such feasibility is by using nonuniformly spaced TDL equalizers (NU-Es), which is the focus of this article. Interested readers can refer to the Further Reading list for references to other solutions that exploit the structure of sparse multipath channels.

3. NONUNIFORMLY SPACED TDL EQUALIZERS (NU-Es)

The distinguishing element of a NU-E is its variable spacings between taps, as opposed to fixed spacings in a uniformly-spaced TDL equalizer (U-E). For all practical purposes, however, a NU-E can be viewed as a U-E with a large number of zero-valued taps, as depicted in Fig. 2, since the spacings of the TDL in a NU-E are usually predetermined by a fixed-rate sampler. In other words, designing a NU-E is equivalent to choosing the best set of tap positions on a fixed-spaced TDL. There are a number

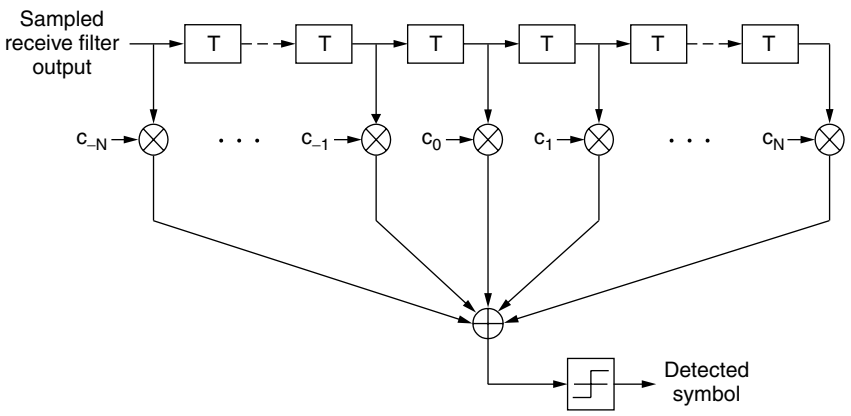


Figure 2. A T-spaced U-LE with $(2N + 1)$ taps. A T-spaced NU-LE can be viewed as a T-spaced U-LE with a large number of zero-valued taps.

of advantages to using a NU-E. First, an appropriately designed NU-E can achieve practically the same performance as a U-E but with fewer taps, hence reducing the computational load of the receiver. This is especially true for equalizing sparse multipath channels, where the number of taps required in a NU-E is proportional to the number of channel taps with large magnitude. In an adaptive NU-E, fewer taps implies faster convergence of the initial learning curve to the optimum tap values, which allows a shorter training sequence to be transmitted. The influence of noisy estimates at the output also tends to lessen without the need to train or track the near-zero-valued taps. For one case with a fractionally spaced TDL in a fixed-point implementation, tracking only a small set of taps decreases the degrees of freedom and thus minimizes the occurrence of the tap wandering phenomenon, an event that arises when tap values deviate from their exact optimum values due to noisy estimates and eventually accumulate to values that are too large and cause overflow.

Despite the many benefits of a NU-E, optimizing its tap positions is not an easy task. Closed-form solution to the optimum tap positions of a finite-length NU-E does not exist in general. Currently, two different approaches are available in the literature for finding these optimum tap positions. The first one, suggested by Raghavan et al. [3], uses a branch-and-bound algorithm to exhaustively search for the best combination of tap positions within a given span of the TDL. The second one, developed by Lee [4], involves numerically solving a set of nonlinear equations to obtain a NU-DFE with tap spacings and tap values that are locally optimum. Note that the normal convention of a fixed-space TDL is removed in the derivation of the nonlinear equations, which means that the resultant tap spacings in the feedforward filter (FFF) can be any real number. Unfortunately, due to the heavy computational burden and the long processing time, both methods are deemed unsuitable for real-time applications, and their usage is mainly restricted to off-line analysis for benchmarking purposes.

To circumvent the difficulty of finding the optimum tap positions, a number of suboptimum tap allocation schemes have also been proposed in the literature. A simple one is the strategy of thresholding, where taps of a U-E are first determined and only those with magnitude above a threshold are retained. Despite its simplicity, this method requires initial training of a large number of taps, which is inefficient. Moreover, the tap values are suboptimum with respect to the retained tap positions, though this problem can easily be corrected by re-optimizing the tap values after thresholding is completed, albeit at a further sacrifice of efficiency. Another easy method is to choose the positions of the channel taps with large magnitude as the positions of a NU-E, as has been adopted by Kocic et al. [5] to obtain a NU-DFE. Other more elaborate solutions include the one by Ariyavisitakul et al. [6], where the FFF of a NU-DFE is designed by selecting the set of taps that maximizes a simplified expression of the output signal-to-noise ratio (SNR), as well as an automated exchange-type algorithm by Lopez et al. [7] that allocates taps to the FFF and the feedback filter (FBF) of a NU-DFE alternatively in an iterative fashion until a desired level of performance is reached.

Similar to the optimum algorithms, the abovementioned suboptimum tap allocation schemes all share the common trait of attempting to find the tap positions of a finite-length NU-E directly. However, this is not the only way to tackle the problem. Given a sparse multipath channel, it turns out that certain infinite-length equalizers are inherently nonuniformly spaced, with the sparseness of those equalizers intimately related to the sparseness of the channel. Consequently, a logical alternative to designing finite-length NU-Es is to first identify the tap positions of an infinite-length NU-E, and then assign a subset of those positions to a finite-length NU-E. This design methodology is originally recognized by Geller et al. [8] and again by Berberidis and Rontogiannis [9], both using the infinite-length zero-forcing (ZF) LEs, and is later extended for designing infinite-length minimum mean-square error (MMSE) LEs and ZF/MMSE-DFEs by Lee and McLane [10]. The beauty of this approach lies in the application of the classical equalization theory to derive the infinite-length NU-Es, thus formally unifying the NU-Es with the well-known U-Es. In addition, provided that an infinite-length equalizer is nonuniformly spaced, the corresponding finite-length NU-E using its tap positions will be asymptotically optimum. Because of these advantages, this methodology is chosen over other existing techniques for designing finite-length NU-Es in this article. Details of the design process are described in the next two sections.

4. INFINITE-LENGTH SYMBOL-SPACED EQUALIZERS FOR SPARSE MULTIPATH CHANNELS

To begin, the infinite-length, symbol-spaced (T-spaced) LE and DFE under the ZF and the MMSE criteria are derived for any given sparse multipath channel. The goal is to determine which of these equalizers are nonuniformly spaced and, if so, their tap positions. Consider the baseband communication system shown in Fig. 3. The data source is an independent and identically distributed (iid) sequence where symbols can be taken from any QAM signaling scheme, and the impulse response of the transmit filter is a square-root raised-cosine (SRRC) pulse. The impulse response of a causal, M-ary sparse multipath channel is given by $h_c(t) = \sum_{i=0}^{M-1} a_i \delta(t - \tau_i T)$, where $\{\tau_i\}$ are restricted to be nonnegative integers

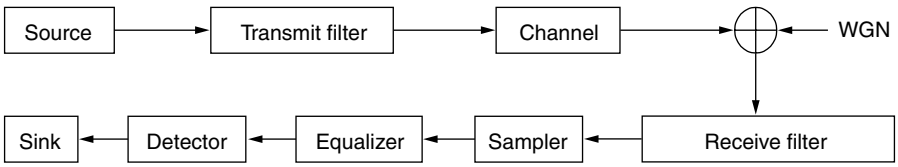


Figure 3. Block diagram of a baseband communication system model. (*Source*: Ref. 10.)

with $\tau_0 = 0$. For brevity, the channel is also denoted as $h_c = [a_0, \ldots, a_{\tau_i}, \ldots, a_{\tau_{M-1}}]$. Complex, zero-mean, white Gaussian noise $n(t)$ with two-sided power spectral density N_0 W/Hz is introduced at the output of the channel, and the noise process is assumed to be independent of the data sequence. At the receiving end, a matched filter (MF) and a SRRC receive filter are used alternatively. With T-spaced sampling at time instants $\{kT\}$, $k = 0, \pm 1, \ldots$, the sampled MF output is $f_s(t) = \sum_{i=-N}^{N} f_i \delta(t - \mu_i T)$, where $N \geq M - 1$ and $\{\mu_i\}$ are integers with $\mu_0 = 0$ and $\mu_i = \mu_{-i}$, and $f_s(t)$ remains a sparse impulse response. If the SRRC receive filter is used, the sampled output is simply $h_c(t)$.

4.1. Linear Equalizers (LEs)

According to classical equalization theory [11,12], the frequency response of the infinite-length, T-spaced ZF-LE under the MF system is

$$C_{\text{ZF-LE,MF}}(w) = \frac{1}{F_s(w)} \tag{1}$$

where $F_s(w) = \sum_{i=-N}^{N} f_i e^{-jw\mu_i T}$ is the (periodic) frequency response of $f_s(t)$. After some simple manipulations, Eq. (1) becomes

$$C_{\text{ZF-LE,MF}}(w) = \frac{k}{1 + \sum_{i=-N(i\neq 0)}^{N} g_i e^{-jw\mu_i T}} \tag{2}$$

where $k = 1/f_0$ and $g_i = k f_i (i \neq 0)$. Let $S(w) = \sum_{i=-N(i\neq 0)}^{N} g_i e^{-jw\mu_i T}$. If $|S(w)| < 1$, Eq. (2) can be expressed as

$$C_{\text{ZF-LE,MF}}(w) = k\left[1 + \sum_{n=1}^{\infty} (-1)^n S^n(w)\right] \tag{3}$$

by making use of the infinite series $1/1 + x = 1 - x + x^2 - x^3 + \cdots$ for $|x| < 1$. Expanding $S^n(w)$ term by term for each n and taking the inverse Fourier transform gives

$$c_{\text{ZF-LE,MF}}(t) = k\left[\delta(t) - \sum_{\substack{i=-N \\ i\neq 0}}^{N} g_i \delta(t - \mu_i T) + \sum_{\substack{i=-N \\ i\neq 0}}^{N} \sum_{\substack{j=-N \\ j\neq 0}}^{N} \times g_i g_j \delta\left(t - (\mu_i + \mu_j)T\right) - \cdots\right] \tag{4}$$

which shows that the tap positions of the ZF-LE are given by the nonnegative-integer-based linear combinations of the multipath delays of $f_s(t)$. Hence, the infinite-length, T-spaced ZF-LE under the MF system is nonuniformly spaced. In the sequel, this derivation method will be called the *infinite-series approach*.

The infinite-length, T-spaced MMSE-LE under the MF system can be determined similarly by using the frequency response

$$C_{\text{MMSE-LE,MF}}(w) = \frac{\sigma_I^2}{\sigma_I^2 F_s(w) + N_0} \tag{5}$$

where σ_I^2 denotes the variance of the data symbols. It is easy to verify that $c_{\text{MMSE-LE,MF}}(t)$ is exactly the same as $c_{\text{ZF-LE,MF}}(t)$ in Eq. (4) except $k = \sigma_I^2/(\sigma_I^2 f_0 + N_0)$, which implies that the ZF-LE and the MMSE-LE share the same tap positions. Hence, the infinite-length, T-spaced MMSE-LE under the MF system is also nonuniformly spaced, and it differs from the ZF-LE only in terms of their tap values. As $N_0 \to 0$, the values of k become the same for both equalizers, and the MMSE-LE converges to the ZF-LE, a well-known result. It is also interesting to note that N_0 actually helps satisfy the condition $|S(w)| < 1$ for the frequency response of the MMSE-LE to be written as an infinite series.

When the SRRC receive filter is used, the frequency response of the infinite-length, T-spaced ZF-LE is

$$C_{\text{ZF-LE,SRRC}}(w) = \frac{1}{H_c(w)} \tag{6}$$

where $H_c(w) = \sum_{i=0}^{M-1} a_i e^{-jw\tau_i T}$ is the frequency response of $h_c(t)$. Let τ_d denote the delay that corresponds to the multipath term with the largest magnitude. Equation (6) can then be expressed as

$$C_{\text{ZF-LE,SRRC}}(w) = \frac{1}{e^{-jw\tau_d T}[a_d + \sum_{i=0(i\neq d)}^{M-1} a_i e^{-jw(\tau_i - \tau_d)T}]} = \frac{k' e^{jw\tau_d T}}{1 + \sum_{i=0(i\neq d)}^{M-1} b_i e^{-jw\nu_i T}} \tag{7}$$

where $k' = 1/a_d$, $b_i = k' a_i$, and $\nu_i = \tau_i - \tau_d$ $(i \neq d)$. Using the infinite-series approach gives

$$c_{\text{ZF-LE,SRRC}}(t) = k'\left[\delta(t + \tau_d T) - \sum_{\substack{i=0 \\ i\neq d}}^{M-1} b_i \delta(t - (\nu_i - \tau_d)T) + \sum_{\substack{i=0 \\ i\neq d}}^{M-1} \sum_{\substack{j=0 \\ j\neq d}}^{M-1} b_i b_j \delta\left(t - (\nu_i + \nu_j - \tau_d)T\right) - \cdots\right] \tag{8}$$

which shows that the infinite-length, T-spaced ZF-LE under the SRRC receive filter system is nonuniformly spaced as well. In essence, its tap positions are obtained by time-shifting the nonnegative-integer-based linear combinations of the time-shifted multipath delays of $h_c(t)$, where the amount of time shift in both instances is τ_d. Note that for the MF system, the multipath term in $f_s(t)$ with

the largest magnitude is always found at $t = 0$. Hence, the time-shifting operations are not required.

For the infinite-length, T-spaced MMSE-LE, its impulse response under the SRRC receive filter system is given by

$$c_{\text{MMSE-LE,SRRC}}(t) = c_{\text{MMSE-LE,MF}}(t) \otimes h_c^*(-t)$$
$$= k \sum_{l=0}^{M-1} a_l^* \left[\delta(t + \tau_l T) - \sum_{\substack{i=-N \\ i\neq 0}}^{N} g_i \delta(t - (\mu_i - \tau_l)T) \right.$$
$$\left. + \sum_{\substack{i=-N \\ i\neq 0}}^{N} \sum_{\substack{j=-N \\ j\neq 0}}^{N} g_i g_j \delta(t - (\mu_i + \mu_j - \tau_l)T) - \cdots \right] \quad (9)$$

where $\otimes$ denotes convolution. This shows that the impulse response of the infinite-length, T-spaced MMSE-LE under the SRRC receive filter system is the sum of M scaled and time-shifted replicas of $c_{\text{MMSE-LE,MF}}(t)$, where the scale factors and delays are determined by $h_c^*(-t)$. In fact, $c_{\text{ZF-LE,SRRC}}(t)$ can also be expressed in the form of (9) with k, $\{g_i\}$ and $\{\mu_i\}$ being the variables of $c_{\text{ZF-LE,MF}}(t)$. Therefore, analogous to the MF system, the tap positions of the ZF-LE and the MMSE-LE under the SRRC receive filter system are identical, and they can be determined from either (8) or (9), even though the two equations appear to be different.

4.2. Decision Feedback Equalizers (DFEs)

To determine the infinite-length, T-spaced ZF-DFE under the MF system, the minimum-phase spectral factorization of $F_s(w)$, namely, $F_s(w) = A_P|P(w)|^2$, where A_P is a constant and $P(w)$ is the monic, causal, and minimum-phase (canonical) factor, is utilized. It is well known that [11] the FFF of the infinite-length, T-spaced ZF-DFE is a whitening filter with frequency response

$$C_{\text{ZF-FFF,MF}}(w) = \frac{1}{A_P P^*(w)} \quad (10)$$

which implies that the tap positions of $c_{\text{ZF-FFF,MF}}(t)$ can be obtained by invoking the infinite-series approach and are the nonnegative-integer-based linear combinations of the multipath delays of $p^*(-t)$. For the FBF, its impulse response is $c_{\text{ZF-FBF,MF}}(t) = p(t) - \delta(t)$, which means that its tap positions are merely the multipath delays of $p(t)$. It is obvious from these two relationships that the sparseness of the infinite-length, T-spaced ZF-DFE is directly dependent on the sparseness of $p(t)$. In fact, $p(t)$ is a sparse impulse response if and only if $h_c(t)$ is sparse and minimum-phase or maximum-phase. If $h_c(t)$ is a mixed-phase channel, $p(t)$ is nonsparse even if $h_c(t)$ is sparse [10]. Therefore, the infinite-length, T-spaced ZF-DFE under the MF system is nonuniformly spaced if and only if $h_c(t)$ is a sparse minimum-phase or maximum-phase channel.

By exploiting the minimum-phase spectral factorization of the received signal-plus-noise spectrum, i.e., $\sigma_I^2 F_s(w) + N_0 = A_Q|Q(w)|^2$, where A_Q is a constant and $Q(w)$ is the canonical factor, the infinite-length, T-spaced MMSE-DFE under the MF system can be derived in the same fashion as for the ZF-DFE. However, because of the noise term N_0, $q(t)$, the inverse Fourier transform of $Q(w)$, is nonsparse for any type of channel in general. As a result, the infinite-length, T-spaced MMSE-DFE under the MF system is uniformly spaced.

When the SRRC receive filter is used, the impulse response of the FFF of the infinite-length, T-spaced ZF-DFE is given by $c_{\text{ZF-FFF,SRRC}}(t) = c_{\text{ZF-FFF,MF}}(t) \otimes h_c^*(-t)$. Hence, similar to the ZF-LE under the SRRC receive filter system, $c_{\text{ZF-FFF,SRRC}}(t)$ is the sum of M scaled and time-shifted replicas of $c_{\text{ZF-FFF,MF}}(t)$ with the scale factors and delays determined by $h_c^*(-t)$. Note that if $h_c(t)$ is a minimum-phase channel, $c_{\text{ZF-FFF,SRRC}}(t)$ should reduce to a scalar. On the other hand, the FBF of the infinite-length, T-spaced ZF-DFE is identical to its counterpart under the MF system: $c_{\text{ZF-FBF,SRRC}}(t) = c_{\text{ZF-FBF,MF}}(t)$. Once again, both relationships hold true for the infinite-length, T-spaced MMSE-DFE. Therefore, conclusions regarding the sparseness of the both DFEs are the same as those under the MF system.

5. FINITE-LENGTH NU-Es FOR SPARSE MULTIPATH CHANNELS

Having derived the various infinite-length, T-spaced equalizers for sparse multipath channels, the next step is to exploit these results for designing finite-length NU-Es. Only the MMSE criterion is considered here, as MMSE equalizers are known to have better performance than ZF equalizers. The system model, notations, and assumptions of Section 4 remain unchanged throughout this section, except that the restriction on the channel will be relaxed later to allow its multipath delays to take on any real number.

For a finite-length, T-spaced NU-LE (i.e., a NU-LE implemented on a T-spaced TDL), its tap positions are obtained directly from its infinite-length counterpart as indicated in Eq. (4) or (8) for the MF system or the SRRC receive filter system, respectively. As a result, as long as the condition for expressing the frequency response of the channel as an infinite-series is valid (i.e., $|S(w)| < 1$), the finite-length, T-spaced NU-LE is asymptotically optimum. Priority is given to the low-order positions, as they correspond to taps with large magnitude. (In the tap allocation algorithms outlined below, this is achieved by choosing small positive values for the coefficients $\{r_i\}$ in determining the nonnegative-integer-based linear combinations.) However, the design procedure for a finite-length, T-spaced NU-DFE is not as straightforward, since its infinite-length counterpart is uniformly spaced and thus provides no useful information on how to select its tap positions. Fortunately, a simple suboptimum strategy can be employed for the MF system, which uses the tap positions on the anticausal side of a NU-LE as the tap positions for the NU-FFF. The NU-FFF under the SRRC receive filter system can then be designed by invoking the time-shifting property on the NU-FFF under the MF system. Once the NU-FFF is fixed, the tap positions of the NU-FBF can be obtained easily by taking the strictly causal portion of the convolution result between the impulse response at the receive filter

output and the impulse response of the NU-FFF. The tap allocation algorithms based on these principles are given below, one for each receive filter system. As an illustration on how to use the algorithms, the tap positions of the NU-LEs for a simple, artificial sparse multipath channel are listed in Table 1.

Algorithm 1 (MF System)

1. To design a finite-length NU-LE:
   a. Identify the positive multipath delays of the impulse response $h(t) = h_c(t) \otimes h_c^*(-t)$ and denote them by the set $\mathcal{S}_{\text{LE}}^1$.
   b. Assign taps at the positions that are the nonnegative-integer-based linear combinations of the elements in $\mathcal{S}_{\text{LE}}^1$ and denote them by $\mathcal{S}_{\text{LE}}^2$. Mathematically, $\mathcal{S}_{\text{LE}}^2 = \{\sum_{i=1}^{s_{\text{num}}} r_i s_i \mid r_i \in \mathcal{Z}^+, s_i \in \mathcal{S}_{\text{LE}}^1, s_{\text{num}} = |S_{\text{LE}}^1|\}$.
   c. Assign taps at the set of positions denoted by $\mathcal{S}_{\text{LE}}^3$, where $\mathcal{S}_{\text{LE}}^3 = \{-s \mid s \in \mathcal{S}_{\text{LE}}^2\}$.
2. To design the FFF of a finite-length NU-DFE:
   a. Assign taps at the set of positions denoted by $\mathcal{S}_{\text{FFF}}$, where $\mathcal{S}_{\text{FFF}} = \mathcal{S}_{\text{LE}}^3$.
3. To design the FBF of a finite-length NU-DFE:
   a. Assign taps at the set of positions denoted by $\mathcal{S}_{\text{FBF}}$, where $\mathcal{S}_{\text{FBF}} = \{s_1 - |s_2| \mid s_1 \in \mathcal{S}_{\text{LE}}^1, s_2 \in \mathcal{S}_{\text{FFF}}, s_1 - |s_2| > 0\}$.

Algorithm 2 (SRRC receive filter system)

1. To design a finite-length NU-LE:
   a. Identify the multipath delays of $h_c(t)$ and denote them by $\mathcal{T}_{\text{LE}}^1$.
   b. Denote as x the delay of $h_c(t)$ that corresponds to the multipath term with the largest magnitude.
   c. Define a new set of positions $\mathcal{T}_{\text{LE}}^2$ by renaming the positions in $\mathcal{T}_{\text{LE}}^1$ relative to x as follows: $\mathcal{T}_{\text{LE}}^2 = \{t - x \mid t \in \mathcal{T}_{\text{LE}}^1\}$.
   d. Define the set $\mathcal{T}_{\text{LE}}^3$ as the nonnegative-integer-based linear combinations of the elements in $\mathcal{T}_{\text{LE}}^2$. Mathematically, $\mathcal{T}_{\text{LE}}^3 = \{\sum_{i=1}^{t_{\text{num}}} r_i t_i \mid r_i \in \mathcal{Z}^+, t_i \in \mathcal{T}_{\text{LE}}^2, t_{\text{num}} = |\mathcal{T}_{\text{LE}}^2|\}$.
   e. Assign taps at the set of positions denoted by $\mathcal{T}_{\text{LE}}^4$, where $\mathcal{T}_{\text{LE}}^4 = \{t - x \mid t \in \mathcal{T}_{\text{LE}}^3\}$.
2. To design the FFF of a finite-length NU-DFE:
   a. Define x as in 1(b). However, if there exists one or more multipath terms with magnitude comparable to that of the largest magnitude term, denote x as the one closest to position 0.
   b. Assign taps at the set of positions denoted by $\mathcal{T}_{\text{FFF}}^1$, where $\mathcal{T}_{\text{FFF}}^1 = \{s - x \mid s \in \mathcal{S}_{\text{FFF}}\}$.
   c. Assign taps at the set of positions denoted by $\mathcal{T}_{\text{FFF}}^2$, where $\mathcal{T}_{\text{FFF}}^2 = \{-t \mid t \in \mathcal{T}_{\text{LE}}^1, t < x\}$.
3. To design the FBF of a finite-length NU-DFE:
   a. Assign taps at the set of positions denoted by $\mathcal{T}_{\text{FBF}}$, where $\mathcal{T}_{\text{FBF}} = \{t_1 - |t_2| \mid t_1 \in \mathcal{T}_{\text{LE}}^1, t_2 \in \mathcal{T}_{\text{FFF}}^1 \cup \mathcal{T}_{\text{FFF}}^2, t_1 - |t_2| > 0\}$.

Table 1. Tap Positions of NU-LEs for Equalizing $\boldsymbol{h_{c_1} = [a_0, a_3, a_5]}$, Where $\boldsymbol{a_0 = 1}$, $\boldsymbol{a_3 = 0.2828 + j0.2828}$, and $\boldsymbol{a_5 = 0.1299 + j0.075}$[a]

13-tap NU-LE (MF)	$\mathcal{S}_{\text{LE}}^2 = \{0, \mathbf{2}, \mathbf{3}, \mathbf{5}, 6, 8, 9\}$; $\mathcal{S}_{\text{LE}}^3 = -\mathcal{S}_{\text{LE}}^2$
6-tap NU-LE (SRRC)	$\mathcal{T}_{\text{LE}}^4 = \{0, \mathbf{3}, \mathbf{5}, 6, 8, 9\}$

[a]The positions in bold are elements of $\mathcal{S}_{\text{LE}}^1$ or $\mathcal{T}_{\text{LE}}^1$, which are used to form the nonnegative-integer-based linear combinations to obtain the other positions.
Source: Ref. 10.

As stated in Section 3, closed-form solution to the optimum tap positions of a finite-length NU-E generally does not exist. However, for a special type of sparse multipath channels whose multipath components are evenly delayed by mT, where m is a positive integer, it can be proved that the only nonzero-valued taps of a finite-length, T-spaced U-LE or U-DFE are located at positions that are multiples of m. In other words, the finite-length, T-spaced U-Es are inherently nonuniformly spaced for such channels and the optimum tap positions are $\{c_0, c_{\pm m}, c_{\pm 2m}, \ldots\}$, which are also the positions assigned by the algorithms. Another feature of the algorithms is their applicability to channels with multipath delays that are nonnegative real numbers. To design a T-spaced NU-E for such channels, an additional step is needed to quantize the assigned tap positions, which are now real numbers, to their nearest integral positions. If a T/2-spaced NU-E is desired, which is usually actually more suitable than a T-spaced NU-E for such channels, the assigned tap positions can be quantized to the nearest positions that are multiples of $\frac{1}{2}$. Similarly, the idea can be extended to design any fractionally spaced equalizers, although fractional spacings smaller than T/2 are seldom used in practice. As an example, Table 2 lists the tap positions of the T- and T/2-spaced NU-DFEs for a sparse multipath channel with delays that are multiples of $\frac{1}{2}$.

While the emphasis has been on the traditional types of NU-Es so far, there exists a different form of NU-DFE, called the *decision-directed feedback equalizer* (NU-DDFE), which has become increasingly popular for equalizing sparse multipath channels [e.g., 13–15] and thus deserves some attention as well. As shown in Fig. 4, the DDFE cancels the postcursor ISI before feedforward filtering via a FBF with impulse response identical to the strictly causal portion of the impulse response of the sampled receive filter output. This feature enables the FBF to fully exploit the sparseness inherent to the output and enjoy a substantial tap reduction, in contrast to the conventional DFE, in which the amount of sparseness in the output is usually diminished after feedforward filtering and thus makes a sizable FBF tap reduction without incurring a significant performance loss difficult. To determine the tap positions of the FFF of a NU-DDFE, Algorithms 1 and 2 can be applied directly, since it has been proved that the FFFs of a DFE and a DDFE are equivalent provided that all postcursor ISI is eliminated [13]. For the FBF of a NU-DDFE, the large magnitude taps in the strictly causal portion of the sampled receive filter output that satisfy a constraint [16] are selected. Note that if a T/2-spaced FFF is employed

Table 2. Tap Positions of NU-DFEs for Equalizing $h_{c_2} = [a_0, a_{1.5}, a_9]$, Where $a_0 = 0.2598 + j0.15$, $a_{1.5} = 1$, and $a_9 = 0.7071 + j0.7071$[a]

(8, 6)-tap NU-DFE (MF)	$\mathcal{S}_{\text{FFF}} = \{0, \mathbf{-1.5}, \mathbf{-7.5}, \mathbf{-9}, -10.5, -15, -16.5, -22.5\}$	T-spaced	$\{0, -1, -2, -7, -8, -9, -15, -16\}$
		T/2-spaced	$\mathcal{S}_{\text{FFF}}$
	$\mathcal{S}_{\text{FBF}} = \{1.5, 6, 7.5, 9\}$	T-spaced T/2-spaced	$\{1, 2, 6, 7, 8, 9\}$
(8, 6)-tap NU-DFE (SRRC)	$\mathcal{T}^1_{\text{FFF}} = \mathcal{S}_{\text{FFF}} - 1.5$	T-spaced	$\{0, -1, -2, -3, -9, -10, -11, -16\}$
	$\mathcal{T}^2_{\text{FFF}} = \{0\}$	T/2-spaced	$\{0, -1.5, -3, -9, -10.5, -16.5, -18, -24\}$
	$\mathcal{T}_{\text{FBF}} = \{1.5, 6, 7.5, 9\}$	T-spaced T/2-spaced	$\{1, 2, 6, 7, 8, 9\}$

[a] A (k_1, k_2)-tap NU-DFE represents one with k_1 FFF taps and k_2 FBF taps.
Source: Ref. 10.

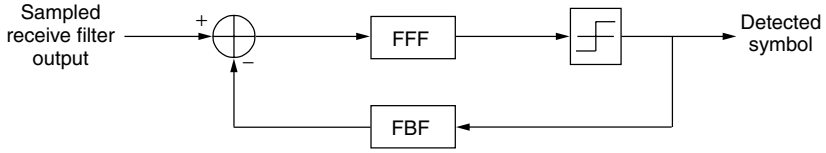


Figure 4. Structure of a DDFE.

in the NU-DDFE, the FBF of the NU-DDFE will be T/2-spaced as well.

Finally, it should be evident that some prior knowledge of the channel, from merely the positions of the taps with large magnitude to a complete estimate of the tap values, is required for most of the existing suboptimum tap allocation schemes discussed in Section 3. For Algorithms 1 and 2, knowing the positions of the taps with large magnitude and which one is the position of the largest magnitude tap are sufficient. Generally speaking, a key property of an appropriate channel estimation algorithm or detection method for the positions of the large magnitude taps is low complexity, so that the overall complexity when added together with a tap allocation scheme and an adaptive NU-E is still lower than the conventional approach of an adaptive U-E without the need of explicit channel estimation. Various channel estimation/detection methods tailored for sparse multipath channels are available in the literature; the more recent ones are specifically customized for use with subsequent equalization. While a discussion of these techniques is outside the scope of this article, a selected few are given in the Further Reading list for the interested reader. It should also be mentioned that, besides implementing an equalizer as an adaptive filter, the equalizer tap values can be computed directly from the channel estimates via fast algorithms with complexity of the order $O(n^2)$, where n is the number of equalizer taps. Unfortunately, such an approach may not be suitable for NU-Es, as these fast algorithms all utilize the Toeplitz nature of the input correlation matrices of U-Es and do not apply to those of NU-Es without the Toeplitz characteristic, which means that only standard algorithms with $O(n^3)$ complexity can be exploited to directly compute the tap values of a NU-E. Therefore, unless the reduction of taps is large, using a NU-E instead of a U-E in this manner may actually increase the computational load.

6. PERFORMANCE EXAMPLE

In this section, a representative example is selected to illustrate some fundamental properties of finite-length NU-Es. The system setup is the same as that shown in Fig. 3. The 4-QAM signaling scheme with constellations $\{\pm 1 \pm j\}$ is chosen and the excess bandwidth of the SRRC transmit and receive filters is set to 35%. Perfect channel knowledge is presumed at the receiver, and optimum tap values of the NU-Es are computed ideally using the matrix inversion approach once their tap positions are determined from the algorithms of Section 5. The bit error rate (BER) after equalization is evaluated analytically through the Beaulieu series as in Ref. 17, and perfect decision feedback is assumed in the DFEs.

Figure 5 shows the performance of the T- and T/2-spaced NU-DFEs for equalizing h_{TV}, one of seven test channels for HDTV systems [2] but with its precursor tap modified from −20 dB to −6 dB relative to the main response to increase the difficulty of equalization, as has been done in Ref. 13. An important result revealed in this plot is that the T-spaced NU-DFE under the MF system only performs better than the one under the SRRC receive filter system at low SNRs, with the crossover point of their BER curves around the SNR of 17.5 dB. This counterintuitive phenomenon is due to the incapability of an equalizer with a limited number of taps, such as a NU-E, to mitigate the extra ISI terms introduced by matched filtering. Although the MF maximizes the SNR before equalization, this gain is only minimal when the influence of noise is insignificant. As a result, if a NU-E is employed at a high SNR region, the benefit of using the MF to maximize the SNR will be insufficient for compensating the detrimental effect of the extra ISI terms introduced, and so the BER suffers. In fact, this phenomenon is even more apparent when the T/2-spaced NU-DFEs of the two systems are compared, with the one under the SRRC receive filter system having a lower BER curve than its counterpart under the MF system for the entire range of SNR shown. This is because a T/2-spaced equalizer can function as a MF, thus allowing the one under the SRRC receive filter system to enjoy the same benefit as the one under the MF system, but without the need to suppress the extra ISI terms that a preceding MF introduces. Therefore, the SRRC receive

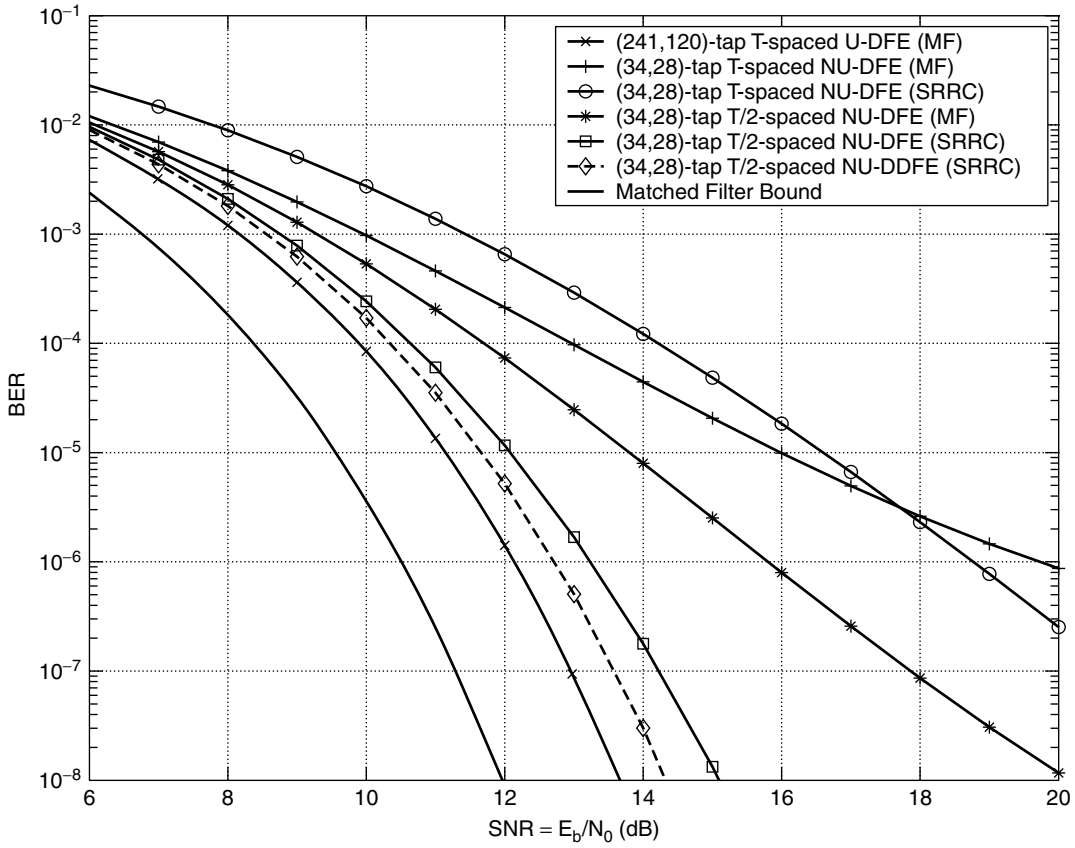


Figure 5. Performance of NU-DFEs for equalizing $h_{\mathrm{TV}} = [a_0, a_{9.68}, a_{10.49}, a_{19.37}, a_{40.35}, a_{106.52}]$, where $a_0 = 0.1549 - j0.4767$, $a_{9.68} = -1$, $a_{10.49} = 0.1$, $a_{19.37} = 0.0389 + j0.1197$, $a_{40.35} = -0.1614 + j0.1173$, and $a_{106.52} = -0.2558 - j0.1859$. (*Source*: Ref. 10.)

filter is a better front-end receive filter than the MF when followed by a NU-E. Note that the performance of the T/2-spaced NU-DDFE under the SRRC receive filter system is also included in Fig. 5. In accordance with the discussion of Section 5, it attains a lower BER curve than its traditional counterpart with only a SNR loss of about 0.7 dB relative to the reference U-DFE, which is used to approximate the performance of the infinite-length equalizer. However, for certain sparse multipath channels, the gain in using a NU-DDFE may not be as significant. More information regarding the strengths and weaknesses of the NU-DDFEs can be found in the paper by Lee and McLane [16].

7. CONCLUSIONS

The ever-increasing data rate in modern communication systems has prompted the quest for new equalization techniques to handle channels with long impulse responses efficiently. NU-Es are suggested in this article as a reduced-complexity solution for equalizing a special family of wireless channels called *sparse multipath channels* that exhibits this undesirable characteristic. Tap positions for the infinite-length, T-spaced ZF/MMSE-LEs/DFEs for such channels can be derived by expressing each frequency response as an infinite series, which in turn lead to simple tap allocation algorithms for designing finite-length NU-Es, including a modified form of the NU-DFE. A fundamental property associated with NU-Es is the nonoptimality of the MF as the front-end receive filter; a better substitute is the SRRC receive filter. Overall, good performance is attained by the NU-Es despite their large reduction in complexity. Together with an appropriate low-complexity channel estimation algorithm, it is conceivable that NU-Es can become a crucial component in future-generation mobile receivers that are expected to operate in a wide variety of channel conditions with ease.

BIOGRAPHIES

Frederick K. H. Lee received the B.Sc. degree in electrical and computer engineering in 1998 from Queen's University, Kingston, Ontario, Canada, where he is currently a Ph.D. candidate. During his postgraduate studies, he has been supported by two postgraduate Scholarships from the Natural Sciences and Engineering Research Council (NSERC) of Canada, a Fessenden Postgraduate Scholarship from the Communications Research Centre (CRC), Canada, and an Ontario Graduate Scholarship. His research interests are communications theory and signal processing algorithms for communications.

Peter McLane has been a Professor at Queen's University since 1969. He is a Fellow of the IEEE and served as the Chairman of the IEEE Communications Society Communication Theory Committee for 3 years. He has served as a Major Project Leader for both Communications and Information Technology Ontario (CITO) and for the Canadian Institute of Telecommunications Research (CITR). He has been a member of both Research and Scholarship Committees with NSERC. He jointly received two research awards: the 1994 Stentor Telecommunications Research Award and the TRIO Feedback Award in 1992. He is one of four authors of the books *Introduction to trellis Coded Modulation with Applications*, originally published by Macmillian in 1991. He has spent academic leaves at UBC, AT&T Bell Labs, Motorola, and Harris Canada.

BIBLIOGRAPHY

1. H. V. Poor and G. W. Wornell, eds., *Wireless Communications: Signal Processing Perspectives*, Prentice-Hall, Upper Saddle River, NJ, 1998.
2. W. F. Schreiber, Advanced television systems for terrestrial broadcasting: Some problems and proposed solutions, *Proc. IEEE* **83**: 958–981 (1995).
3. S. A. Raghavan, J. K. Wolf, L. B. Milstein, and L. C. Barbosa, Nonuniformly spaced tapped-delay-line equalizers, *IEEE Trans. Commun.* **COM-41**: 1290–1295 (1993).
4. I. Lee, Optimization of tap spacings for the tapped delay line decision feedback equalizer, *IEEE Commun. Lett.* **COMML-5**: 429–431 (2001).
5. M. Kocic, D. Brady, and M. Stojanovic, Sparse equalization for real-time digital underwater acoustic communications, *Proc. IEEE OCEANS'95*, 1995, pp. 1417–1422.
6. S. Ariyavisitakul, N. R. Sollenberger, and L. J. Greenstein, Tap-selectable decision-feedback equalization, *IEEE Trans. Commun.* **COM-45**: 1497–1500 (1997).
7. M. J. Lopez and A. C. Singer, A DFE coefficient placement algorithm for sparse reverberant channels, *IEEE Trans. Commun.* **COM-49**: 1334–1338 (2001).
8. B. Geller et al., Equalizer for video rate transmission in multipath underwater communications, *IEEE J. Ocean. Eng.* **OE-21**: 150–155 (1996).
9. K. Berberidis and A. A. Rontogiannis, Efficient decision feedback equalizer for sparse multipath channels, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, ICASSP'00*, 2000, pp. 2725–2728.
10. F. K. H. Lee and P. J. McLane, Design of nonuniformly-spaced tapped-delay-line equalizers for sparse multipath channels, *IEEE Trans. Commun.* (in press); see also *Proc. IEEE Global Telecommunications Conf., GLOBECOM'01*, 2001, pp. 1336–1343.
11. E. A. Lee and D. G. Messerschmitt, *Digital Communication*, 2nd ed., Kluwer, Boston, MA, 1994.
12. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.
13. I. J. Fevrier, S. B. Gelfand, and M. P. Fitz, Reduced complexity decision feedback equalization for multipath channels with large delay spreads, *IEEE Trans. Commun.* **COM-47**: 927–937 (1999).
14. P. De, J. Bao, and T. Poon, A calculation-efficient algorithm for decision feedback equalizers, *IEEE Trans. Consumer Electron.* **CE-45**: 526–532 (1999).
15. M. Stojanovic, L. Freitag, and M. Johnson, Channel-estimation-based adaptive equalization of underwater acoustic signals, *Proc. IEEE OCEANS'99*, 1999, pp. 985–990.
16. F. K. H. Lee and P. J. McLane, Comparison of two nonuniformly-spaced decision feedback equalizers for sparse multipath channels, *Proc. IEEE Int. Conf. Commun. ICC'02*, 2002, pp. 1923–1928.
17. J. E. Smee and N. C. Beaulieu, Error-rate evaluation of linear equalization and decision feedback equalization with error propagation, *IEEE Trans. Commun.* **COM-46**: 656–665 (1998).

FURTHER READING

Alternative Equalization Techniques for Sparse Multipath Channels

N. Benvenuto and R. Marchesani, The Viterbi algorithm for sparse channels, *IEEE Trans. Commun.* **COM-44**: 287–289 (1996).

N. C. McGinty, R. A. Kennedy, and P. Hoeher, Parallel trellis Viterbi algorithm for sparse channels, *IEEE Commun. Lett.* **COMML-2**: 143–145 (1998).

R. Cusani and J. Mattila, Equalization of digital radio channels with large multipath delay for cellular land mobile applications, *IEEE Trans. Commun.* **COM-47**: 348–351 (1999).

S. Chowdhury, M. D. Zoltowski, and J. S. Goldstein, Structured MMSE equalization for synchronous CDMA with sparse multipath channels, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, ICASSP'01*, 2001, pp. 2113–2116.

K. Chugg, A. Anastasopoulos, and X. Chen, *Iterative Detection: Adaptivity, Complexity Reduction and Applications*, Kluwer Academic Publishers, 2001, Chapter 3.

Selected Channel Estimation/Detection Methods for Sparse Multipath Channels

Y. F. Cheng and D. M. Etter, Analysis of an adaptive technique for modeling sparse systems, *IEEE Acoust. Speech Signal Process.* **ASSP-37**: 254–264 (1989).

M. Kocic and D. Brady, Complexity-constrained RLS estimation for sparse systems, *Proc. Conf. Information Science Systems, CISS'94*, 1994, pp. 420–425.

J. Homer, I. Mareels, R. R. Bitmead, B. Wahlberg, and F. Gustafsson, LMS estimation via structural detection, *IEEE Trans. Signal Process.* **SP-46**: 2651–2663 (1998).

I. Kang, M. P. Fitz, and S. B. Gelfand, Blind estimation of multipath channel parameters: A modal analysis approach, *IEEE Trans. Commun.* **COM-47**: 1140–1150 (1999).

I. Ghauri and D. T. M. Slock, Structured estimation of sparse channels in quasi-synchronous DS-CDMA, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, ICASSP'00*, 2000, pp. 2873–2876.

Y. Jin and B. Friedlander, On the performance of equalizers for sparse channels, *Proc. Asilomar Conf. Signals, Syst., Comput.* **34**: 1757–1761 (2000).

W. Sung, D. J. Shin, and I. K. Kim, Maximum-likelihood tap selection for equalization of land mobile radio channels, *Proc. IEEE Global Telecommunications Conf., GLOBECOM'01*, 2000, pp. 3326–3330.

S. F. Cotter and B. D. Rao, Sparse channel estimation via matching pursuit with application to equalization, *IEEE Trans. Commun.* **COM-50**: 374–377 (2002).

O

OPTICAL COUPLERS

GERD KEISER
PhotonicsComm Solutions, Inc.
Newton Center, Massachusetts

1. INTRODUCTION

Optical couplers play a key role in wavelength-division multiplexing (WDM) applications for combining and separating wavelength channels, tapping off power for monitoring purposes, or adding and dropping specific wavelengths at a particular point in an optical fiber communication link [1–3]. Most optical couplers are passive devices in the sense that they do not need to be powered externally to perform their function on optical signals. Fundamentally, optical couplers connect three or more fibers to combine, split, or redirect light signals.

Since optical couplers perform many different functions, they can be made in several configurations, as shown in Fig. 1. The T coupler, Y coupler, or 1×2 coupler is a three-port device that is mainly used to tap off a portion of the light from a throughput fiber into a second fiber. The relative optical power level in each output branch is usually given in percentages. The design can be tailored to achieve any coupling ratio between the two outputs. This coupler nominally is used for signal-monitoring applications. In this case, a tradeoff between coupling loss in the primary fiber and an adequate level of power required for the measurement threshold in the secondary branch shows that a 10% tap is the optimal configuration [4]. This means that 90% of the input optical power continues through the device and 10% is tapped off for signal monitoring purposes.

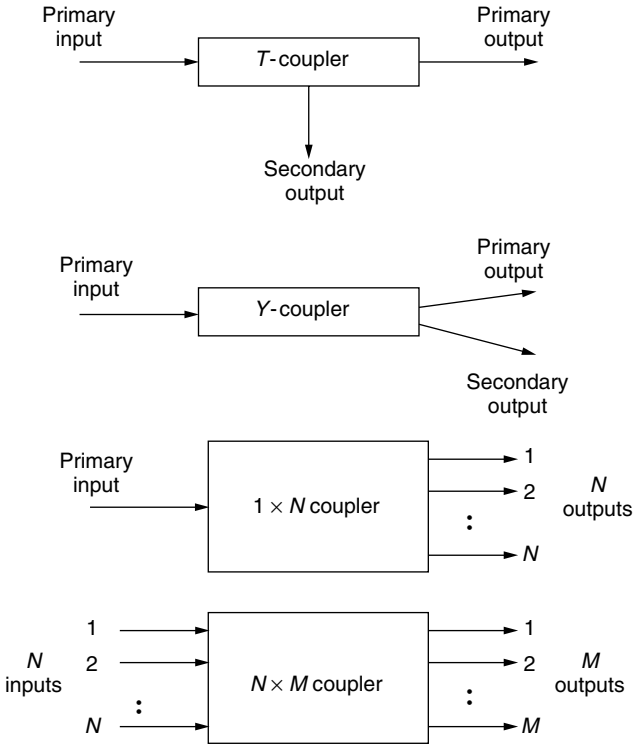


Figure 1. Example configurations for optical couplers.

The $1 \times N$ or tree coupler has one input fiber and N output fibers. In the most general case, this device is not wavelength dependent and it divides all the input optical power equally among the N output ports. Many of these devices are directional, which means that their function depends on the direction in which the light passes through it.

A more general configuration is the $N \times M$ or star coupler, which has N input ports and M output ports. In the broadest application, star couplers combine the light streams from two or more input fibers and divide them among several output fibers. In the general case, the splitting is done uniformly for all wavelengths, so that each of the M outputs receives $1/M$ of the power entering the device. A common fabrication method for an $N \times N$ coupler is to fuse together the cores of N single-mode fibers over a length of a few millimeters. The optical power inserted through one of the N fiber entrance ports gets divided uniformly into the cores of the N output fibers through evanescent power coupling in the fused region.

Wavelength-selective couplers form a more versatile category of devices for WDM applications. Among the technologies used for making these devices are 2×2 fused-fiber couplers, coupled planar waveguides, Mach–Zehnder interferometers, fiber Bragg gratings, and phased-array waveguide gratings.

2. COUPLER CONSTRUCTIONS

The 2×2 coupler is a simple fundamental device that we will use here to demonstrate the operational principles. These devices can be fabricated by microoptic, optical fiber, or integrated optic methods. A common construction is the fused-fiber coupler [3,5–7]. This is fabricated by twisting together, melting, and pulling two single-mode fibers so they get fused together over a uniform section of length W, as shown in Fig. 2. Each input and output fiber has a long tapered section of length L, since the transverse dimensions are gradually reduced down to that of the coupling region when the fibers are pulled during the fusion process. The total draw length is $2L + W$. This device is known as a fused biconical tapered coupler. Here P_0 is the input power, P_1 is the throughput power, and P_2 is the power coupled into the second fiber. The parameters P_3 and P_4 are extremely low signal levels (-50 to -70 dB below the input level) resulting from backward reflections and scattering due to bending in and packaging of the device.

As the input light P_0 propagates along the taper in fiber 1 and into the coupling region, an increasingly larger portion of the input field now propagates outside the core

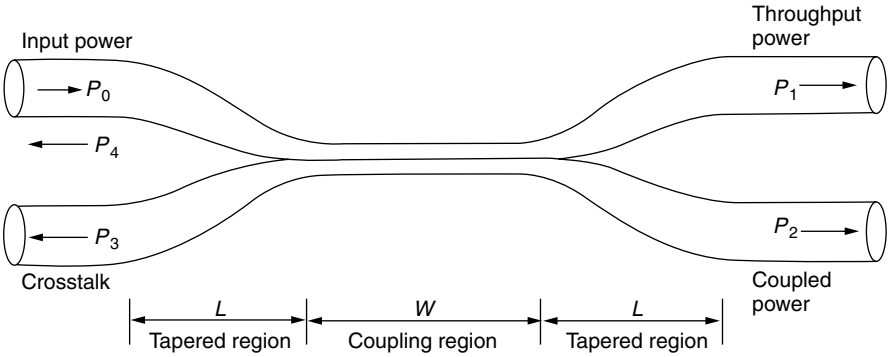


Figure 2. Cross-sectional view of a fused-fiber coupler having a coupling region W and two tapered regions of length L. The total span $2L + W$ is the coupler draw length.

of the input fiber and is coupled into the adjacent fiber. Depending on the dimensioning of the coupling region, any desired fraction of this decoupled field can be coupled into fiber 2. By making the tapers very gradual, only a negligible fraction of the incoming optical power is reflected back into either of the input ports. Thus these devices are also known as directional couplers.

Fused-fiber couplers have an intrinsic wavelength dependence in the coupling region. The optical power coupled from one fiber to another at a specific wavelength can be varied through three parameters: the axial length of the coupling region over which the fields from the two fibers interact; the size of the reduced radius r in the coupling region; and Δr, the difference in radii of the two fibers in the coupling region. In making a fused-fiber coupler, the coupling length W is normally fixed by the width of the heating flame that is used in the melting process, so that only L and r change as the coupler is elongated. Typical values for W and L are a few millimeters, the exact values depending on the coupling ratios desired for a specific wavelength, and $\Delta r/r$ is around 0.015. Assuming that the coupler is lossless, the expression for the power P_2 coupled from one fiber to another over an axial distance z is

$$P_2 = P_0 \sin^2(\kappa z) \tag{1}$$

where κ is the *coupling coefficient* describing the interaction between the fields in the two fibers. By conservation of power, for identical-core fibers we have

$$P_1 = P_0 - P_2 = P_0[1 - \sin^2(\kappa z)] = P_0 \cos^2(\kappa z) \tag{2}$$

Wavelength-dependent multiplexers can also be made using Mach–Zehnder interferometry techniques [8]. Figure 3 illustrates the constituents of an individual

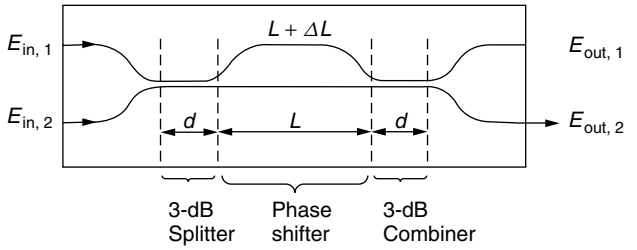


Figure 3. Constituents of an individual Mach–Zehnder interferometer (MZI). E_{in} and E_{out} are electric field intensities.

Mach–Zehnder interferometer (MZI). This 2×2 MZI consists of three stages: an initial 3-dB directional coupler which splits the input signals, a central section where one of the waveguides is longer by ΔL to give a wavelength-dependent phase shift between the two arms, and another 3-dB coupler which recombines the signals at the output. The function of this arrangement is that, by splitting the input beam and introducing a phase shift in one of the paths, the recombined signals will interfere constructively at one output and destructively at the other. The signals then finally emerge from only one output port.

A grating is an important element in WDM systems for combining and separating individual wavelengths. Basically a grating is a periodic structure or perturbation in a material. This variation in the material has the property of reflecting or transmitting light in a certain direction depending on the wavelength. Thus gratings can be categorized as either transmitting or reflecting gratings.

Figure 4 shows a simple concept of a demultiplexing function using a fiber Bragg grating [9,10]. To extract the desired wavelength, a circulator is used in conjunction with the grating. In a three-port circulator, an input signal on one port exits at the next port. For example,

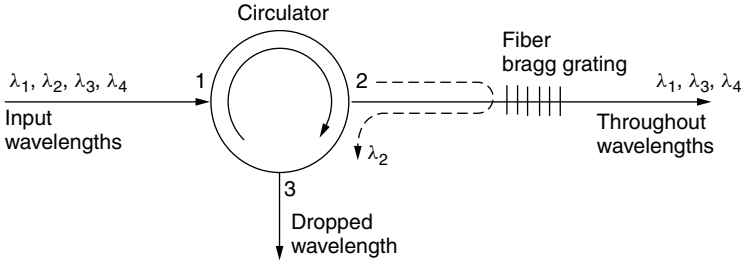


Figure 4. A simple wavelength demultiplexing function using a fiber Bragg grating.

an input signal at port 1 is sent out at port 2. Here the circulator takes the four wavelengths entering port 1 and sends them out port 2. All wavelengths except λ_2 pass through the grating. Since λ_2 satisfies the Bragg condition of the grating, it is reflected, enters port 2 of the circulator, and exits at port 3. More complex multiplexing and demultiplexing structures with several gratings and several circulators can be realized with this scheme.

A highly versatile WDM device is based on using an arrayed waveguide grating. This device can function as a multiplexer, a demultiplexer, a drop-and-insert element, or a wavelength router. A variety of design concepts have been examined [11,12]. The arrayed waveguide grating is a generalization of the 2×2 Mach–Zehnder interferometer multiplexer. One popular design consists of M_{in} input and M_{out} output slab waveguides and two identical focusing planar star couplers connected by N uncoupled waveguides with a propagation constant β. The lengths of adjacent waveguides in the central region differ by a constant value ΔL, so that they form a Mach–Zehnder-type grating, as Fig. 5 shows. For a pure multiplexer, we can take $M_{\text{in}} = N$ and $M_{\text{out}} = 1$. The reverse holds for a demultiplexer, that is $M_{\text{in}} = 1$ and $M_{\text{out}} = N$. In the case of a network routing application, we can have $M_{\text{in}} = M_{\text{out}} = N$.

3. PERFORMANCE CHARACTERISTICS

In specifying the performance of an optical coupler, one usually indicates the percentage division of optical power between the output ports by means of the splitting ratio or coupling ratio. Referring to Fig. 2, where P_0 is the input power and P_1 and P_2 the output powers, then

$$\text{Splitting ratio} = \left(\frac{P_2}{P_1 + P_2}\right) \times 100\% \tag{3}$$

By adjusting the parameters so that power is divided evenly, with half of the input power going to each output, one creates a 3-dB coupler. A coupler could also be made in which, for example, almost all the optical power at 1500 nm goes to one port and almost all the energy around 1300 nm goes to the other port.

In the analysis above, we have assumed for simplicity that the device is lossless. However, in any practical coupler there is always some light that is lost when a signal goes through it. The two basic losses are excess loss and insertion loss. The excess loss is defined as the ratio

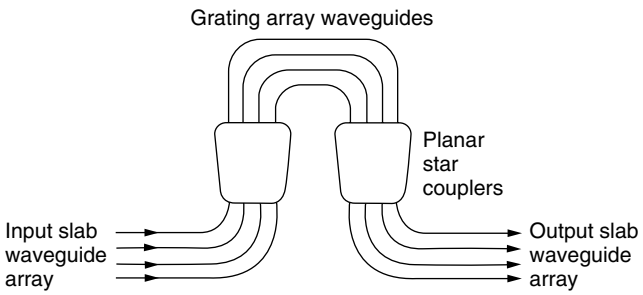


Figure 5. Adjacent waveguides in the central region differ in length to form a Mach–Zehnder-type grating.

of the input power to the total output power. Thus, in decibels, the excess loss for a 2×2 coupler is

$$\text{Excess loss} = 10 \log \left(\frac{P_0}{P_1 + P_2}\right) \tag{4}$$

The insertion loss refers to the loss for a particular port-to-port path. For example, for the path from input port i to output port j, we have, in decibels:

$$\text{Insertion loss} = 10 \log \left(\frac{P_i}{P_j}\right) \tag{5}$$

Another performance parameter is crosstalk, which measures the degree of isolation between the input at one port and the optical power scattered or reflected back into the other input port. That is, it is a measure of the optical power level P_3 shown in Fig. 2:

$$\text{Crosstalk} = 10 \log \left(\frac{P_3}{P_0}\right) \tag{6}$$

The principal role of any star coupler is to combine the powers from N inputs and divide them equally among M output ports. Techniques for creating star couplers include fused fibers, gratings, microoptic technologies, and integrated optics schemes. The fused-fiber technique has been a popular construction method for $N \times N$ star couplers. For example, 7×7 devices and 1×19 splitters or combiners with excess losses at 1300 nm of 0.4 and 0.85 dB, respectively, have been demonstrated. However, large-scale fabrication of these devices for $N > 2$ is limited because of the difficulty in controlling the coupling response between the numerous fibers during the heating–pulling process.

In an ideal star coupler the optical power from any input is evenly divided among the output ports. The total loss of the device consists of its splitting loss plus the excess loss in each path through the star. The splitting loss is given in decibels by

$$\text{Splitting loss} = -10 \log \left(\frac{1}{N}\right) = 10 \log N \tag{7}$$

For a single input power P_{in} and N output powers, the excess loss in decibels is given by

$$\text{Fiber star excess loss} = 10 \log \left(\frac{P_{\text{in}}}{\sum_{i=1}^{N} P_{\text{out},i}}\right) \tag{8}$$

The insertion loss and crosstalk can be found from Eqs. (5) and (6), respectively.

An alternative is to construct star couplers by cascading 3-dB couplers. Figure 6 shows an example for an 8×8 device formed by using twelve 2×2 couplers. This device could be made from either fused-fiber or integrated-optic components. As can be seen from this figure, a fraction $1/N$ of the launched power from each input port appears at all output ports. A limitation to the flexibility or modularity

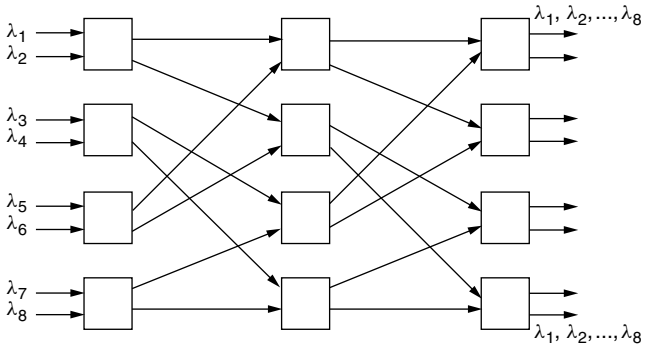


Figure 6. Example for an 8×8 device formed by using twelve 2×2 couplers.

of this technique is that N is a multiple of 2, that is, $N = 2^n$ with the integer $n \geq 1$. The consequence is that if an extra node needs to be added to a fully connected $N \times N$ network, the $N \times N$ star needs to be replaced by a $2N \times 2N$ star, thereby leaving $2(N-1)$ new ports being unused. Alternatively, one extra 2×2 coupler can be used at a port with the result that the outputs of the two new ports have an additional 3-dB loss.

As can be deduced from Fig. 6, the number of 3-dB couplers needed to construct an $N \times N$ star is

$$N_c = \frac{N}{2}\log_2 N = \frac{N}{2}\frac{\log N}{\log 2} \tag{9}$$

since there are $N/2$ elements in the vertical direction and $\log_2 N = \log N / \log 2$ elements horizontally.

If the fraction of power traversing each 3-dB coupler element is F_T, with $0 \leq F_T \leq 1$ (i.e., a fraction $1 - F_T$ of power is lost in each 2×2 element), then the excess loss in decibels is

$$\text{Excess loss} = -10\log(F_T^{\log_2 N}) \tag{10}$$

The splitting loss for this star is again given by Eq. (7). Thus, the total loss experienced by a signal as it passes through the $\log_2 N$ stages of the $N \times N$ star and gets divided into N outputs is, in decibels,

$$\begin{aligned}\text{Total loss} &= \text{splitting loss} + \text{excess loss} \\ &= -10\log\left(\frac{F_T^{\log_2 N}}{N}\right) = -10\left(\frac{\log N \log F_T}{\log 2} - \log N\right) \\ &= 10(1 - 3.322\log F_T)\log N \end{aligned} \tag{11}$$

This shows that the loss increases logarithmically with N.

BIOGRAPHY

Gerd Keiser is the founder and president of Photonics-Comm Solutions, Inc., Newton Center, Massachusetts, a firm specializing in consulting and education for the optical communications industry. He has 25 years experience at Honeywell, GTE, and General Dynamics in designing and analyzing telecommunication components, links, and networks. He is the author of the books *Optical Fiber Communications* (3rd ed. 2000) and *Local Area Networks* (2nd ed. 2002) published by McGraw-Hill. Dr. Keiser is an IEEE fellow and received GTE's prestigious Leslie Warner Award for work in ATM switch development. He earned his B.A. and M.S. degrees in mathematics and physics from the University of Wisconsin and a Ph.D. in solid state physics from Northeastern University, Boston, Massachusetts.

BIBLIOGRAPHY

1. G. Keiser, *Optical Fiber Communications*, 3rd ed., McGraw-Hill, Burr Ridge, IL, 2000, Chap. 10.
2. J. Hecht, *Understanding Fiber Optics*, 4th ed., Prentice-Hall, Upper Saddle River, NJ, 2002, Chap. 15.
3. V. J. Tekippe, Passive fiber optic components made by the fused biconical taper process, *Fiber Integr. Opt.* **9**(2): 97–123 (1990).
4. M. Hoover, New coupler applications in today's telephony networks, *Lightwave* **17**: 134–140 (March 2000) (see *http://www.light-wave.com*).
5. A. Ankiewicz, A. W. Snyder, and X.-H. Zheng, Coupling between parallel optical fiber cores — critical examination, *J. Lightwave Technol.* **4**: 1317–1323 (Sept. 1986).
6. E. Pennings, G.-D. Khoe, M. K. Smit, and T. Staring, Integrated-optic versus micro optic devices for fiber-optic telecommunication systems: A comparison, *IEEE J. Select. Top. Quant. Electron* **2**: 151–164 (June 1996).
7. R. W. C. Vance and J. D. Love, Back reflection from fused biconic couplers, *J. Lightwave Technol.* **13**: 2282–2289 (Nov. 1995).
8. R. Syms and J. Cozens, *Optical Guided Waves and Devices*, McGraw-Hill, New York, 1992.
9. Y. Fujii, High-isolation polarization-independent optical circulator coupled with single-mode fibers, *J. Lightwave Technol.* **9**: 456–460 (April 1991).
10. R. Ramaswami and K. N. Sivarajan, *Optical Networks*, 2nd ed., Morgan Kaufmann, San Francisco, 2002.
11. M. K. Smit and C. van Dam, PHASAR-based WDM devices: Principles, design and applications, *IEEE J. Select. Top. Quant. Electron* **2**: 236–250 (June 1996).
12. H. Takahashi, K. Oda, H. Toba, and Y. Inoue, Transmission characteristics of arrayed waveguide $N \times N$ wavelength multiplexers, *J. Lightwave Technol.* **13**: 447–455 (March 1995).

OPTICAL CROSSCONNECTS

Li Fan
OMM, Inc.
San Diego, California

1. INTRODUCTION

Massive information demand in the Internet is creating enormous needs for the capacity and communication bandwidth expansion in the service providers and the carriers. With the expectation that the new data traffic will have

exponential growth, the service providers are under pressure to find a new technology, which can dramatically reduce the cost of hardware and network management as well as solving the bandwidth bottleneck. Instead of chasing the growing bandwidth, service providers are desperate to keep ahead of the competition. In particular, they want to boost capacity by orders of magnitude. Optical networks are digital communications systems that use light waves in the fiber as a medium for the transmission or switching of data. This new optical layer is the future to providing cost-effective capacity. Dense wavelength division multiplexing (DWDM) is the ideal solution to dramatically increase bandwidth. With the enormous channels and traffic, optical crossconnects (OXC) is the emerging technology and component that will deliver and manage this new optical layer and the services running on it. Optical crossconnects can be used for protection and restoration in optical networks, bandwidth provisioning, wavelength routing, and network performance monitoring. It is one of the key elements for routing optical signals in an optical network or system for long-haul communication, metro area, and local access add-drop as shown in Fig. 1.

2. OPTICAL CROSSCONNECTS FABRIC

Currently, carrier backbones already carry information traffic with light over fiber. For long-distance fiber, the light has to be converted back into electrical and periodically regenerated depending on the fiber type. Most current OXC in fact use an electronic core for switching such as the synchronous optical network (SONET) system, fiber distributed data interface (FDDI) switches, asynchronous transfer mode (ATM) switches, and ethernet switches with fiber optic interfaces to convert the photons into electrical signals in order to switch them from one fiber to another at junction points. Sometimes the electronic switch is referred to as optical-electrical-optical (OEO) switch as shown in Fig. 2. In this OEO crossconnects, the input/output signals are optical, but receivers convert the input signals to electrical signals, then use electronic components to route the channels through the core. At the transceiver module, the electrical signals are converted back into photons. This solution is not future proof since when the data rate increases, the expensive transceivers and the electrical switch core have to be replaced.

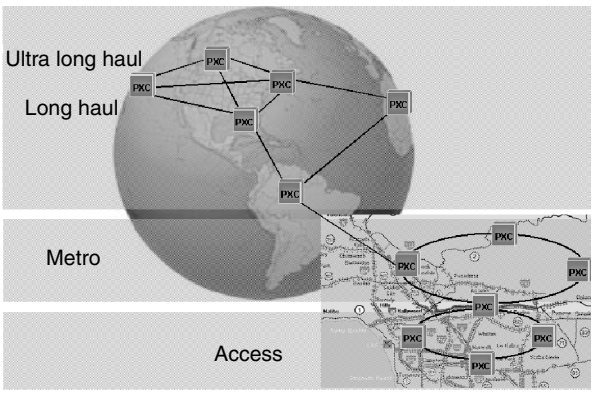


Figure 1. Use optical crossconnects to manage the optical network for wavelength routing, network performance monitoring, bandwidth provisioning, protection, and restoration in optical networks.

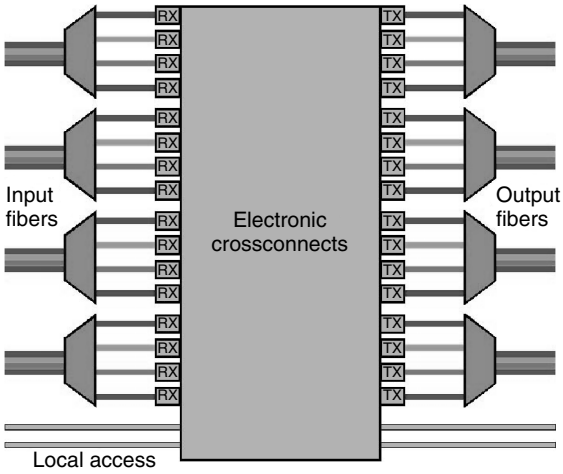


Figure 2. OEO electronic crossconnects. The wavelengths from DWDM fibers are converted into electrical signals by receivers. The switching and routing are done by the electronic core. The output signals are converted back into optical by transmitters.

All-optical crossconnects use light waves exclusively from end to end. The data is maintained in original optical format from the input fiber to the switch element. The data format is unchanged from switching elements to the output fiber. Sometimes the all-optical crossconnects is referred to as OOO crossconnects as shown in Fig. 3, which stands for optical-optical-optical. It is preferred to use the terminology "photonic crossconnects" for all-optical crossconnects rather than "optical crossconnects." This is to designate the fact that the data path of the switch is purely photonic, with no electrical conversions. The all-optical crossconnects are much more attractive because of the avoidance of the conversion stages and because the core switch is independent of data rate and data protocol, making the

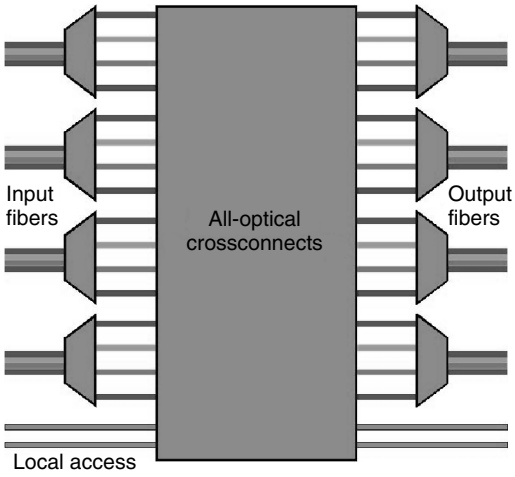


Figure 3. All-optical crossconnects. The data is maintained in original optical format from input fiber through the core switch to the output fiber. There is no need for expensive and high-power consumption high-speed electronics and transmitters and receivers.

cross-connect ready for future data rate upgrades. It provides a valuable capability for handling high bandwidth traffic, as there is no need for expensive and high-power consumption high-speed electronics, transmitters, and receivers. The system becomes less expensive in addition to the reduction of complexity. The all-optical crossconnects improve reliability and reduce the footprint when compared with OEO solutions. Another major benefit of all-optical devices is their greater scalability over OEOs.

Some advantages of the all-optical crossconnects are also disadvantages when we try to coexist this technology with current network. All-optical crossconnects maintain the original signal from input to output fibers without signal regeneration for cost saving. They use erbium doped fiber amplifiers (EDFA) to boost the signal, not regeneration. However, this approach also loses the advantages of signal regeneration. The network design would be challenging to route the same wavelength from the source to the destination and through the entire multirings or meshes network, eliminating the transponders and removing the capability of wavelength conversion. There is no visibility of bit error rate (BER) or monitoring. An all-optical network only has lambda (wavelength) level granularity and cannot perform sub-lambda mixing and grooming.

To combine the scalability of all-optical crossconnects and the wavelength regeneration and grooming, a compromise design is to integrate the all-optical crossconnects ports with an OEO switch as shown in Fig. 4. The majority of the fabric is a fully connected all-optical crossconnects. A small percent of the output ports and input ports are integrated with local add/drop switching through and OEO router. Since any input port can be connected to any output port and go through the OEO router, the data stream of the selected channel can be detected and processed by software at the individual packet level; the wavelength is capable of 3R (reshape/retime/regenerate) and wavelength translation if the network design is sufficiently advanced.

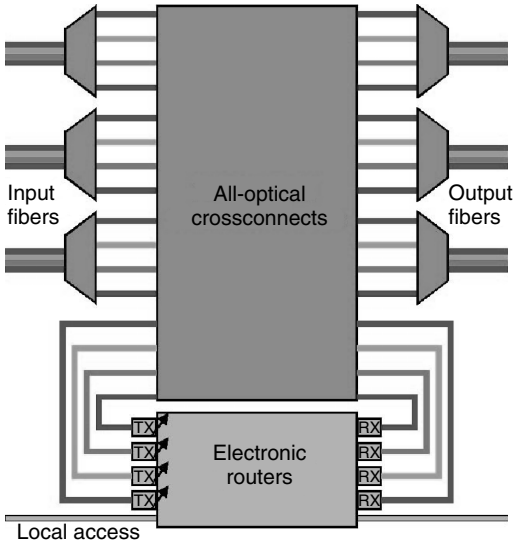


Figure 4. Compromised network design integrated all-optical crossconnects with electronic routers for 3R regeneration and wavelength conversion.

The more complex system may also use additional optical repeaters at each I/O module for 3R regeneration. The repeater is a receiver that can receive any wavelength, directly connected to a transmitter whose output wavelength matches the channel of the wavelength multiplexer.

3. OPTICAL CROSSCONNECTS ARCHITECTURE: 2D AND 3D

The architectures of OXC can be categorized into two approaches. The first configuration is the 2D approach. The devices are arranged in a crossbar configuration. Figure 5 shows the schematic diagram of an 8×8 crossconnects with switches arranged in a 2D array. Each mirror has a digital movement. The switch has only either on or off status, which makes the driving scheme very straightforward. The device can be controlled with a simple TTL signal and does not require feedback control. When a switch is set at the off position, the optical signal will pass through the switch with minimum insertion loss. When the switch is activated, it will bounce the optical signal by 90° and direct the light to the output fiber. The reflection can be achieved by total internal reflection from different refractive index or it can be reflected by a free-space micromirror. Additional functionality can be achieved for adding or dropping optical signals if plane 3 and plane 4 are utilized. Because of the crossbar arrangement, the required mirror number is not linear to port count. The mirror number is equal to N^2. This approach is ideal for small port count crossconnects. However, a large port count crossconnects such as 64×64 will require 4096 switching elements, which is still challenging for current technology.

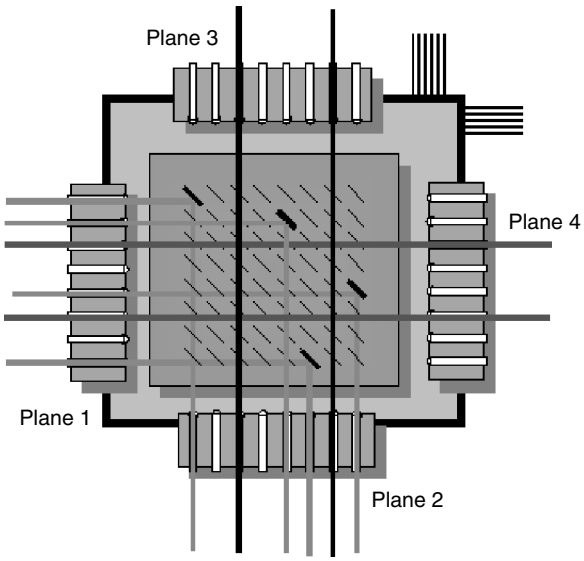


Figure 5. Schematic diagram of 2D digital crossconnects. It needs N^2 switches to configure an $N \times N$ crossconnects. Each switch has only two states: on and off. Optical signal will pass through the switch with minimum lose when the switch is at off position. The light will be reflected at 90° when the switch is activated.

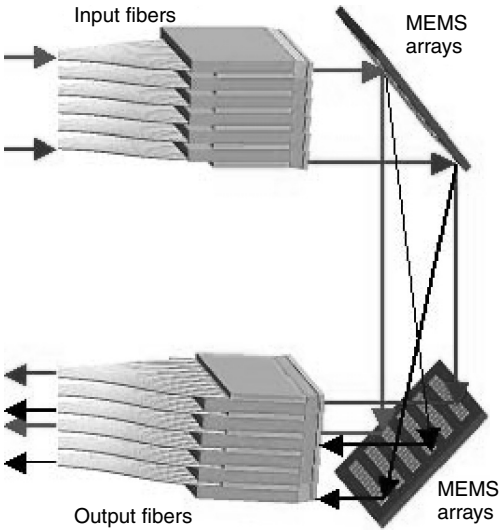


Figure 6. Schematic diagram of 3D crossconnects. The switch number is linearly proportional to channel number. It needs 2N switches to construct $N \times N$ crossconnects. Each mirror has N different states, which are able to point at all the output mirrors with high precision.

The second architecture for the crossconnects is the 3D approach as shown in Fig. 6. This approach has the advantage for switch number (2N), which is linearly proportional to channel number. The connection path is established by tilting two analog mirrors and directing the light from the input fiber to the output fiber. Each mirror has N different states, which need to be well controlled. The 3D approach is the most promising architecture for very large port counts, more than hundred or thousand. The analog fine-tuning of the mirror angle is another unique functionality from 3D that can be used to minimize the insertion loss and equalize uniformity. The drawback of this approach is the control loop and monitoring which can be very complex. The mirror stability, shock/vibration, and long-term drift need to be carefully controlled. The overall total states number for 2D and 3D are the same (Fig. 7). They both need $2N^2$ different states to complete the switching functionality. However, 2D puts more pressure on the switch number; 3D has a smaller switch number but puts more burden on mirror angle control.

Usually OXC are desired to be nonblocking; any input fiber can be switched to any output fiber. With 16 input fibers and 40 wavelengths, the crossconnects size can easily grow to several hundred or even thousand port count at the first installation. It is putting an incredible burden on current technology, especially a few years ago when there were only 2×2 switches commercially available. Even though the large port count crossconnects can be built from 2×2 switching elements, it is not practical or cost effective especially for the performance and scalability. Large nonblocking networks can be constructed with smaller switch fabric by using a multistage Clos network to reduce the number of crosspoints compared to simple matrices [1]. In blocking crossconnects, some connections cannot be established for certain choices or the switch paths are limited to certain zone area. However, the blocking switches can be used as an advantage to reduce the complexity of the crossconnects and enable a larger port count system from smaller modules. For example the wavelength-selective crossconnects (WSXC) is a stack of $N \times N$ switches, each dedicated to signals of the same wavelength as drawn in Fig. 8. For a network with sixteen input fibers, each carrying forty wavelengths, the crossconnects would need to be 640×640 if the system needs to be nonblocking. Because of the lack of wavelength conversion in the fabric, the wavelength is not interchangeable between different wavelengths. Therefore, crossconnects are needed only among the same wavelength. The same functionality network can be built with 40 packages of 16×16 switches, a pay-as-you-grow business model using smaller switches as building elements. New fabrics are added when new wavelengths are turned on. The total bandwidth capacity of a WSXC can be extremely large with low first-installed cost and scalability.

4. TECHNOLOGY

4.1. Planar Lightwave Circuit: Thermo-Optic and Electro-Optic

The planar lightwave circuits (PLC) are constructed with rectangular cross sections of different refraction index materials. The section that transmits the light has a slightly higher refractive index, so that total internal reflection acts to guide the light within the waveguides. The key elements of PLC switches are two directional couplers and the Mach-Zehnder interferometer (MZI). A directional coupler consists of two waveguides very close to each other, so that light waves can be coupled from one to the

	Array size	Switch number	States for each switch	Total states
2D	$N \times N$	N^2	2	$N^2 \times 2 = 2N^2$
3D	$N \times N$	2N	N	$2N \times N = 2N^2$

Figure 7. Two different approaches for OXC show different trade-off. However, the total states for each configuration are identical. Both 2D and 3D require total number of $2N^2$ different states in order to complete $N \times N$ crossconnects functionality. 2D approach has more weight on switch number, which will show bottleneck with larger port count. 3D approach only requires switch number linear proportional to channel number. However, the burden is shifted to the complexity of the switch design.

Figure 8. Wavelength-Selective Crossconnects uses small $N \times N$ switches as building block. The small switch port count (N) is equal to the input fiber number. The total package number (M) is increased when new wavelengths are turned on. The total bandwidth capacity of a WSXC is N times M, which can be extremely large with scalability and low first-install cost.

other. The MZI is a pair of waveguides with identical path lengths. They are separated far enough and will not couple energy between these two waveguides. The incoming light from the input waveguide is spit 50/50 in the first directional coupler. The upper branch goes through a controlled path while the lower branch goes through a reference pathway as illustrated in Fig. 9. The refractive index can be thermally controlled by locally heating the thin film heater above the waveguide [2], or it can be controlled by electro-optic lithium niobate technology [3]. Since the controlled branch and the reference branch of the MZI have identical path length, the light energy will be recombined in the second directional coupler and switch to the upper branch when the controller is off. The controller will change the refractive index and effectively change the path length and phase of the upper branch. The interference will switch the light wave to the lower branch when the controller is on.

Waveguide technology was among the first all-optical switches to be developed, typically in the 1×1, 1×2, and 2×2 range. Because of the planar technology, larger crossconnects can be formed by integrating basic 2×2 components on the same wafer. The optical performance parameters such as crosstalk and insertion loss could be unacceptable for optical network application. However, this technology is capable of integrating variable optical attenuators (VOA) optical switch and wavelength selective elements on the same substrate. It does not require free-space collimator alignment.

4.2. Microfluid

The microfluid and microbubble utilize the interfaces of different refractive index and cause total internal reflection to redirect the light beam. Bubble switches demonstrated the switching mechanism from intersecting waveguides [4]. At each intersection, a trench is etched into the waveguide. The trenches are filled with an index-matching fluid to direct the light in the through states. To direct the light, the thermal inkjet-like matrix-controller silicon chip element heats up the fluid and creates a microbubble in the liquid. The location of the bubble is at the intersection between the input waveguide and output waveguide. The light is reflected by total internal reflection from the liquid/bubble interface as shown in Fig. 10.

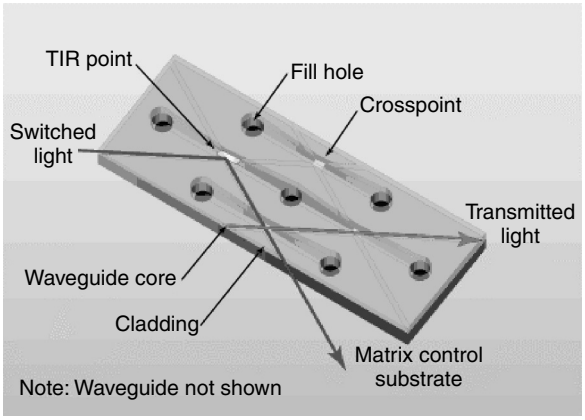


Figure 10. Bubble switch directs light from an input to an output, by using a thermal inkjet element to create a bubble in an index-matching fluid at the intersection between the input waveguide and the desired output waveguide. The light is redirected by means of total internal reflection. Courtesy of Agilent Technologies Inc.

A thermal-capillarity optical switch utilizes a similar total internal reflection concept [5]. The switch element consists of an upper substrate and an intersecting waveguide substrate that has a slit at each crossing point with refractive index matching oil in it and a pair of micro-heaters that produce a thermal gradient along the slit as shown in Fig. 11. The matching oil within the slit is driven by a decrease in interfacial tension of the air-oil interface caused by thermo-capillarity. This switch element also has bi-stable self-latching achieved by capillary pressure that depends on the slit width.

4.3. Liquid Crystal

Liquid crystal crossconnects uses liquid crystal to rotate the optical beam polarization by applying electric voltage to adjust the molecules orientation. Based on this rotation, a beam steering router displaces the signal to one of

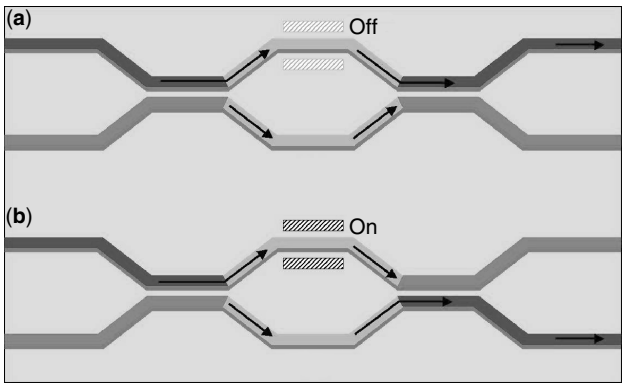


Figure 9. Light from input fiber enters a waveguide at the edge of the optical wafer and goes through a 50/50 split in a directional coupler. One branch goes through a refractive index controlled path (by thermal-optic or electro-optic) while the other goes through a reference pathway.

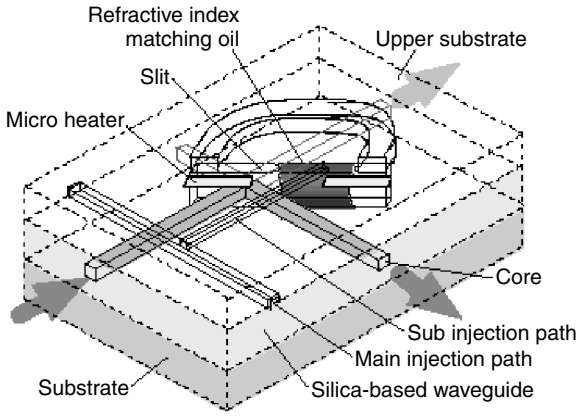


Figure 11. The basic structure of the thermo-capillarity optical switch element. The matching oil within the slit is driven by a decrease in interfacial tension of the air-oil interface caused by thermo-capillarity. Courtesy of NTT Electronics Corp.

two possible paths. Most vendors use this technology for variable optical attenuators rather than switches. This is because liquid crystals can be used to adjust the amount of light that passes through them, rather than simply deflect it. In the switch market, the technology is probably best suited to mid-size wavelength-selective devices. The absence of moving parts and low-power consumption make them a good candidate for test and measurement applications.

4.4. MEMS Micro Mirror

Micro-Electro-Mechanical Systems (MEMS) have been deployed for over a decade in a number of applications such as airbag sensors, projection systems, scanners, and microfluidics. Continued technical developments in the last five years has extended MEMS applications to include optical networking with devices such as all-optical switching fabrics and variable optical attenuators. The MEMS technology has opened up many new possibilities for free-space optical systems. The first commercial MEMS photonic crossconnects were made available in 1999. MEMS technology is using a batch-fabrication process, which is a similar process for making large scale integrated (VLSI) circuits. Wafer scale and chip scale batch process produced MEMS components with high-precision controlled movements. The micro-mechanical structures are smaller, lighter, faster, and more cost effective compared to traditional macro-scale components. The MEMS has become a very good candidate for optical applications, which require stringent reliability, precision, performance, and scalability. The dimension of the micromirror ranges from 0.1 mm to 1 mm for the "sweet spot" of OXC design space. The performance will be strongly limited by Gaussian beam diffraction if the dimension is too small. The large port count crossconnects will not be compatible or scalable with IC processes if the unit switch dimension is too large. The shock/vibration stability and the speed will also miss the SONET specs and Telcordia requirements. Figure 12 shows the OMM 2D digital mirror design and the 16×16 array by surface micromachining technology. The mirror is actuated by a simple TTL signal. The device is not sensitive to driving voltage fluctuations. Since a large force can be generated from this type of gap-closing actuator, the mechanical structure can be built more robust compared to a 3D scanning device. The MEMS array and collimators are hermetically sealed in the package. Maximum insertion loss as low as 1.7 db and 3.1 dB have been obtained for 8×8 and 16×16 2D crossconnects [6].

In the trend of building larger port-count systems and larger mirrors, the optics designs prefer to have a flat mirror surface. Bulk micromachining constructs MEMS devices from single crystal silicon. This technology also enables the possibility of using vertical comb drive to actuate the mirror with large force and more linear response. Figure 13 shows the Lucent LambdaRouter all-optical switch based on Bell Labs's innovative MicroStar 3D MEMS technology [7]. Each mirror angle can be continuously adjusted by voltage.

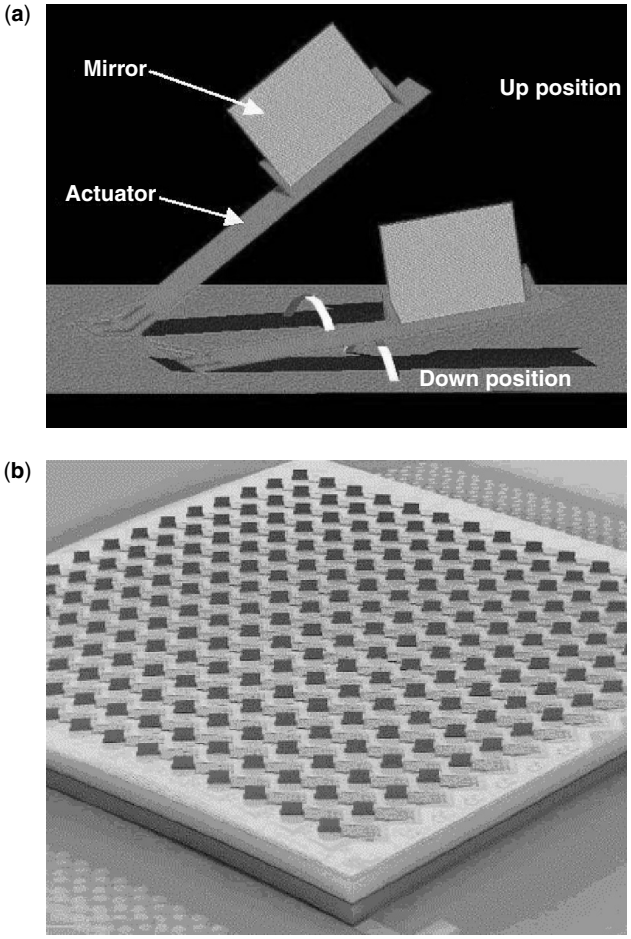


Figure 12. Digital mirrors design from OMM Inc.: (**a**) schematic of basic mirror/switch element. The mirror has bi-stable positions actuated by electrostatic force. (**b**) SEM image of a 16×16 crossconnects with fully populated 256 digital mirrors.

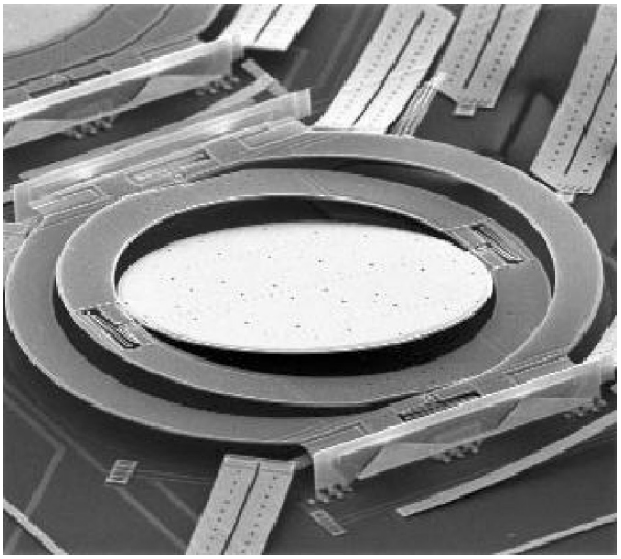


Figure 13. Lucent LambdaRouter based on 3D scanning mirror technology. Each gimbal-structural mirror is able to scan along X and Y axes simultaneously.

5. SUMMARY

Despite the overall slowdown in telecommunications since the year 2000, MEMS based crossconnects have proven to be a reliable and promising technology that can fulfill the stringent requirements of telecommunication industries. Optical crossconnects allow the carriers to add new flexibility and scalability to their optical network. 2D crossconnects are ideal for small port count, less than 64 channels. 3D crossconnects will be the candidate for large-scale networks such as a hundred or thousand channels. It is still not clear what is the optimized module size or approach for the future optical network. It is also not clear how the all-optical network is going to merge with existing opaque systems. The 2D array size has grown from 4×4, 8×8 to 16×16 and 32×32. There is high possibility that the 2D array could achieve larger size with new technology and design improvements. There is also potential that 3D crossconnects will lower the cost, which is competitive with 2D technology. The optical MEMS has gone through an explosive development period. It evolved from the first concept of a fixed stand-up mirror five years ago. Right now it has become a well-accepted technology for fiber communication. There is certainly development space for crossconnects once the demand is generated from the industry.

BIOGRAPHY

Li Fan is cofounder and chief technologist at OMM and is responsible for MEMS design and technology development. He has also successfully commercialized photonic crossconnect ranges from 4×4 to 32×32 based on his MEMS design. He received the Ph.D. in electrical engineering from UCLA in 1998. His research included Optical-Micro-Electro-Mechanical Systems (OMEMS), self-assembly micro-XYZ stage, fiber optical cross-connect, and beam-steering vertical cavity surface-emitting lasers (VCSEL).

BIBLIOGRAPHY

1. C. Clos, A study of nonblocking switching networks, *Bell Syst. Tech. J.* **32**: 406–424 (March 1953).
2. T. Goh et al., Low-loss and high-extinction-ratio silica-based strictly nonblocking 16×16 thermooptic matrix switch, *IEEE Photon. Technol. Lett.* **10**: 810–812 (June 1998).
3. E. J. Murphy, "Photonic switching," I. P. Kaminow and T. L. Koch, eds. *Optical Fiber Telecommunications III B*, Academic Press, New York, 1997, pp. 463–501.
4. J. E. Fouquet, Compact optical cross-connect switch based on total internal reflection in a fluid-containing planar lightwave circuit, *in Proc. Optical Fiber Communication* (*OFC*), TuM1, (2000).
5. M. Makihara, M. Sato, F. Shimokawa, and Y. Nishida, Micromechanical optical switches based on thermocapillary integrated in waveguide substrate, *J. Lightwave Tech.* **17**: 14–18 (1999).
6. P. D. Dobbelaere et al., Digital MEMS for optical switching, *IEEE Comm. Mag.* 88–95 (March 2002).
7. V. Aksyuk et al., Low insertion loss packaged and fiber-connectorized Si surface-micromachined reflective optical switch, *in Proc. Solid-State Sensor and Actuator Workshop*, Hilton Head Island, SC, June 1998, pp. 79–82.

OPTICAL FIBER COMMUNICATIONS

GERD KEISER
PhotonicsComm Solutions, Inc.
Newton Center, Massachusetts

1. INTRODUCTION

1.1. Overview

A major need in human society is the desire to send messages from one distant place to another. Some of the earliest communication systems were based on optical signaling. For example, in the eighth century B.C. the Greeks used a fire signal to send alarms, call for help, or announce certain events. Since then, many forms of communication methodologies have appeared. The basic motivation behind each new form was either to improve the transmission fidelity, to increase the data rate so that more information can be sent, to increase the transmission distance between relay stations, or a combination of these factors. The basic trend in these improvements was to move to higher and higher frequencies of the electromagnetic spectrum. The reason for this is that, in electrical systems, information is usually transferred over the communication channel by superimposing the data onto a sinusoidally varying electromagnetic wave, which is known as the *carrier*. Since the amount of information that can be transmitted is directly related to the frequency range over which the carrier operates, increasing the carrier frequency in turn increases the available transmission bandwidth, and, consequently, provides a larger information capacity.

Although these transmission links generally made use of radio, microwave, and copper-wire technologies, there has always been an interest in using light to communicate [1–4]. The reason for this is that, in addition to the optical fiber's inherently wide bandwidth capability, its dielectric nature renders it immune to electromagnetic interference and offers excellent electrical isolation, particularly in electrically hazardous environments. In addition, its low weight and hair-sized dimensions offer a distinct advantage over large, heavy copper cables, which is important not only for saving space in underground and indoor ducts but also for reducing the size and weight of cables on aircraft and in ships. Kao and Hockman first proposed the use of low-loss glass fiber in 1966, when they suggested that the intrinsic loss of silica-based glass could be made low enough to enable its use as a guiding channel for light [5]. The fabrication of a low-loss optical fiber by researchers at Corning in 1970 provided the key technology for finally realizing this in a practical way [6].

The optical spectrum ranges from about 50 nm (ultraviolet) to about 100 μm (far infrared), the visible region being the 400–700-nm band. Optical fiber communication systems operate in the 800–1600-nm wavelength

band. In optical systems it is customary to specify the band of interest in terms of wavelength, instead of frequency as in the radio region. However, with the advent of high-speed multiple-wavelength systems in the mid-1990s, researchers began specifying the output of optical sources in terms of optical frequency. The reason for this is that in optical sources such as mode-locked semiconductor lasers, it is easier to control the frequency of the output light, rather than the wavelength, in order to tune the device to different emission regions. Of course, the different optical frequencies ν are related to the wavelengths λ through the fundamental equation $c = \nu\lambda$. Thus, for example, a 1552.5-nm wavelength light signal has a frequency of 193.1 THz (193.1×10^{12} Hz).

1.2. Optical Fiber Link Applications

Communication networks composed of optical fiber links are sometimes referred to as *lightwave* or *photonic* systems. Network architectures using multiple wavelength channels per optical fiber can be utilized in local-area, metropolitan-area, or wide-area applications to connect hundreds or thousands of users having a wide range of transmission capacities and speeds. The use of multiple wavelengths greatly increases the capacity, configuration flexibility, and growth potential of this backbone. Moderate-speed regional networks attached to this backbone provide applications such as interconnection of telephone switching centers, access to satellite transmission facilities, and access to mobile-phone base stations. More localized, lower-speed networks offer a wide variety of applications such as telephony services to homes and businesses, distance learning, Internet access, CATV (cable television), security surveillance, and electronic mail (email). A major motivation for developing these sophisticated networks has been the rapid proliferation of information exchange desired by institutions such as commerce, finance, education, health, government, security, and entertainment. The potential for this information exchange arose from the ever-increasing power of computers and data-storage devices.

Once researchers showed in 1970 that it was possible to make low-loss fibers, the optical fiber communication field expanded rapidly to provide a broadband medium for transporting voice, video, and data traffic. In fact, optical fiber technology has been a key factor contributing to the extraordinary growth of global telecommunications. Optical fiber was being installed worldwide at the rate of 4800 km per hour by the year 2000, which is equivalent to a cable-laying rate of three time around the world every day [7,8]. Along with this high installation rate come numerous technological advances in photonic components. These advances permit more and more wavelengths to be transmitted at ever-increasing speeds on an individual optical fiber, which is resulting in an annual two-fold increase in the data-carrying capacity of an individual fiber strand. Table 1 illustrates this with a few of the many installations that have taken place since 1980. As shown in that table, in the year 2000, commercial systems were capable of transmitting 400 Gbps (gigabits per second) over distances of 640 km without regenerating the signal. To put this in perspective, this is equivalent to sending 12,000 encyclopedic volumes every second.

1.3. Basic Link Elements

Figure 1 shows typical components that are found within an optical fiber link. The key sections are a transmitter consisting of a light source and its associated drive circuitry, a cable offering mechanical and environmental protection to the optical fibers contained inside, and a receiver consisting of a photodetector plus amplification and signal-restoring circuitry. Additional components include optical amplifiers, connectors, splices, couplers, and regenerators (for restoring the signal shape characteristics). The cabled fiber is one of the most important elements in an optical fiber link. In addition to protecting the glass fibers during installation and service, it may contain copper wires for powering optical amplifiers or signal regenerators, which are needed periodically in long-distance links for amplifying and reshaping the signal.

Analogous to copper cables, the installation of optical fiber cables can be either aerial, in underground or indoor ducts, undersea, or buried directly in the ground. As a result of installation and/or manufacturing limitations, individual cable lengths will range from several hundred meters to several kilometers for terrestrial links. Cable lengths for oceanic links can be several tens of kilometers. Practical considerations such as reel size and cable weight determine the actual length of a single cable section. The shorter segments tend to be used when the cables are pulled through ducts. Longer lengths are used in aerial,

Table 1. Examples of Types of Multimode (MM) and Single-Mode (SM) Optical Fiber Systems Installed Since 1980

Year	Fiber Type	Wavelength (nm)	WDM Channels	Bit Rate per Channel	Bit Rate per Fiber	Regenerator Spans (km)
1980	MM	820	1	45 Mbps	45 Mbps	7
1985	SM	1300	1	417 Mbps	417 Mbps	50
1987	SM	1300	1	1.7 Gbps	1.7 Gbps	50
1992	SM	1300	1	2.5 Gbps	2.5 Gbps	50
1995	SM	1550	8	2.5 Gbps	20 Gbps	360
1997	SM	1550	16	2.5 Gbps	40 Gbps	360
1999	SM	1550	80	2.5 Gbps	200 Gbps	640
1999	SM	1550	40	10 Gbps	400 Gbps	640
2000	SM	1550	80	10 Gbps	800 Gbps	500

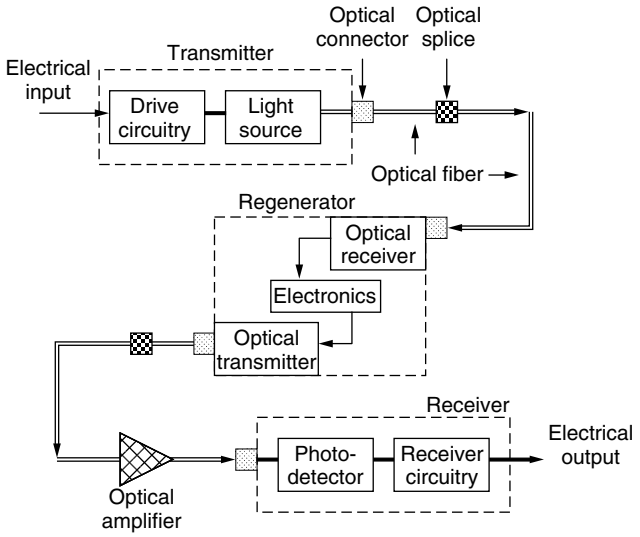


Figure 1. Typical components found within an optical fiber link.

directly buried, or undersea applications. Splicing together individual cable sections forms continuous transmission lines for these long-distance links. For undersea installations, the splicing and repeater-installation functions are carried out on board a specially designed cable-laying ship [9].

An optical fiber nominally is a thin cylindrical strand of two layers of glass surrounded by an elastic buffer coating, as shown in Fig. 2. The central cylinder has a radius a and an index of refraction n_1. This cylinder is known as the *core* of the fiber. The core is surrounded by a solid glass *cladding*, which has a refractive index n_2 that is slightly less than n_1. Since the core refractive index is larger than the cladding index, electromagnetic energy at optical frequencies can propagate along the fiber core through internal reflection at the core–cladding interface. Single-mode fibers, which sustain a single propagating mode along the core, have nominal core diameters of 8–12 μm.

One of the principal characteristics of an optical fiber is its attenuation as a function of wavelength, as shown in Fig. 3. Early technology made exclusive use of the 800–900-nm wavelength band, since in this region the fibers made at that time exhibited a local minimum in the attenuation curve, and optical sources and photodetectors operating at these wavelengths were available. This region is referred to as the *first window*. The large attenuation spikes in early fibers were due to absorption by water molecules (hydroxyl ions) in the glass. By reducing the concentration of hydroxyl ions and metallic impurities in the fiber material, in the 1980s manufacturers were able to fabricate optical fibers with very low loss in the 1100–1600-nm region. This spectral band is referred to

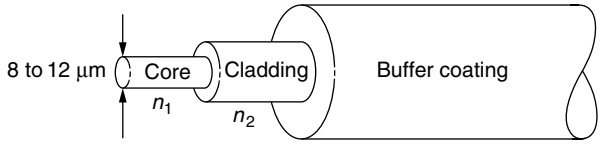


Figure 2. Physical configuration of an optical fiber.

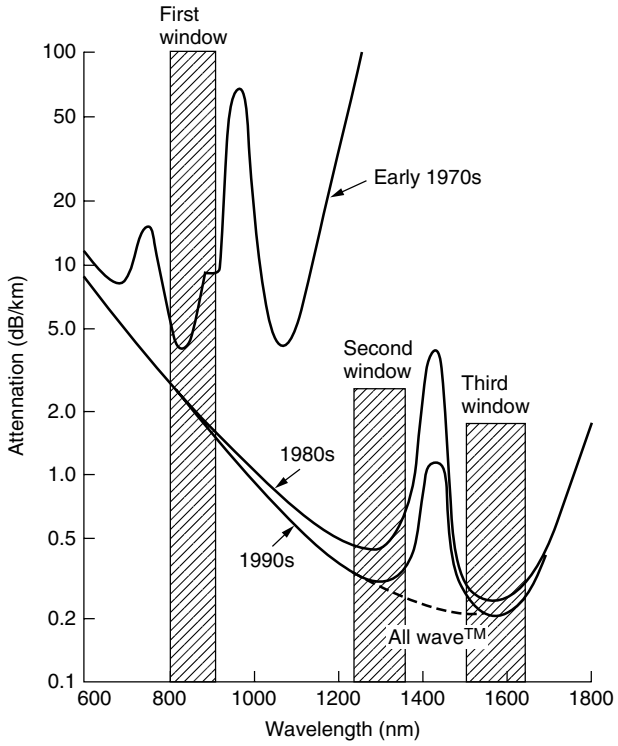


Figure 3. Attenuation as a function of wavelength.

as the *long-wavelength region*. As Fig. 3 shows, the two windows defined here are called the *second window* centered around 1310 nm and the *third window*, centered around 1550 nm.

In 1998 a new ultra-high-purification process patented by Lucent Technologies eliminated virtually all water molecules from the glass fiber material. By dramatically reducing the water attenuation peak around 1400 nm, this process opened the transmission region between the second and third windows to provide around 100 nm more bandwidth than in conventional single-mode fibers, as shown by the dashed line in Fig. 3. This particular AllWave fiber, which was specifically designed for metropolitan networks, gave local service providers the ability to cost-effectively deliver up to hundreds of optical wavelengths simultaneously.

Once the cable is installed, a light source that is dimensionally compatible with the fiber core is used to launch optical power into the fiber. Semiconductor light-emitting diodes (LEDs) and laser diodes are suitable for this purpose, since their light output can be modulated rapidly by simply varying the bias current at the desired transmission rate, thereby producing an optical signal. The electric input signals to the transmitter circuitry for the optical source can be either of an analog or digital form. For high-rate systems (usually greater than 1 Gbps), direct modulation of the source can lead to unacceptable signal distortion. In this case, an external modulator is used to vary the amplitude of a continuous light output from a laser diode source. In the 800–900-nm region the light sources are generally alloys of GaAlAs. At longer wavelengths (1100–1600 nm) an InGaAsP alloy is the principal optical source material.

After an optical signal is launched into a fiber, it will become progressively attenuated and distorted with increasing distance because of scattering, absorption, and dispersion mechanisms in the glass material. At the receiver a photodiode will detect the weakened optical signal emerging from the fiber end and convert it to an electric current (referred to as a *photocurrent*). Silicon photodiodes are used in the 800–900-nm region. The primary photodiode material in the 1100–1600-nm region is an InGaAs alloy.

The design of an optical receiver is inherently more complex than that of the transmitter, since it has to interpret the content of the weakened and degraded signal received by the photodetector. The principal figure of merit for a receiver is the maximum optical power necessary at the desired data rate to attain either a given error probability for digital systems or a specified signal-to-noise ratio for an analog system. The ability of a receiver to achieve a certain performance level depends on the photodetector type, the effects of noise in the system, and the characteristics of the successive amplification stages in the receiver.

1.4. Wavelength-Division Multiplexing

An interesting and powerful aspect of an optical communication link is that many different wavelengths can be sent along a fiber simultaneously in the 1300–1600-nm spectrum. The technology of combining a number of wavelengths onto the same fiber is known as *wavelength-division multiplexing* (WDM). Figure 4 shows the basic WDM concept [10,11]. Here N independent optically formatted information streams, each transmitted at a different wavelength, are combined with an optical multiplexer and sent over the same fiber. Note that each of these streams could be at a different data rate. Each information stream maintains its individual data rate after being multiplexed with the other streams, and still operates at its unique wavelength. Conceptually, the WDM scheme is the same as frequency-division multiplexing (FDM) used in microwave radio and satellite systems. WDM systems in which the channels are closely spaced are referred to as *dense WDM* (DWDM) systems.

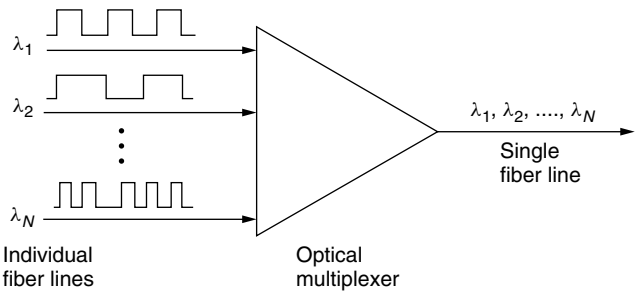


Figure 4. The basic WDM concept.

To realize this WDM scheme, one needs specialized components to efficiently multiplex an aggregate of wavelength channels into an optical fiber at one point and to divide them into their original individual channels at another location. Other WDM components are used to selectively add or drop one or more channels at specific points along a fiber link. These components include fiber Bragg gratings, arrayed waveguide gratings, dielectric thin-film interference filters, acoustooptic tunable filters, and Mach–Zehnder filters.

Figure 5 shows a simple concept of a demultiplexing function using a fiber Bragg grating. To extract the desired wavelength, a *circulator* is used in conjunction with the grating. In a three-port circulator, an input signal on one port exits at the next port. For example, an input signal at port 1 is sent out at port 2. Here, the circulator takes the four wavelengths entering port 1 and sends them out at port 2. All wavelengths except λ_2 pass through the grating. Since λ_2 satisfies the Bragg condition of the grating, it is reflected, enters port 2 of the circulator, and exits at port 3. More complex multiplexing and demultiplexing structures with several gratings and several circulators can be realized with this scheme.

1.5. Optical Amplifiers

Traditionally, when setting up an optical link, one formulates a power budget and adds repeaters when the path loss exceeds the available power margin. To amplify an optical signal with a conventional repeater, one performs photon-to-electron conversion, electrical amplification, retiming, pulseshaping, and then electron-to-photon conversion. Although this process works well for moderate-speed single-wavelength operation, it can be fairly complex and expensive for high-speed multiwavelength systems. Thus, a great deal of effort has been expended to develop all-optical amplifiers for the 1300-nm and the 1550-nm long-wavelength transmission windows of optical fibers. The main ones in use are *erbium-doped fiber amplifiers* (EDFAs) and *Raman fiber amplifiers* [12,13]. An EDFA can amplify optical signals in the 1530–1610-nm range, whereas Raman fiber amplifiers can amplify signals from 1270 to 1670 nm.

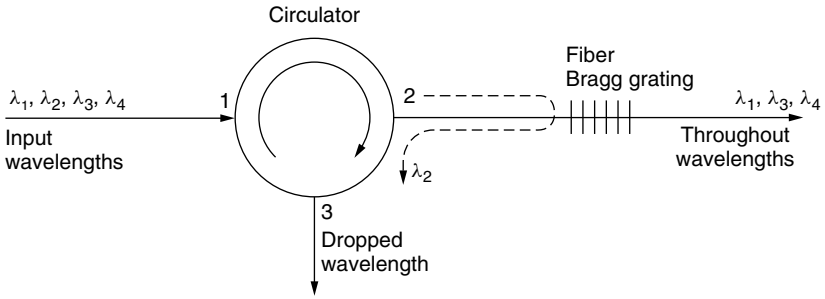


Figure 5. A simple wavelength demultiplexing function using a fiber Bragg grating.

Optical amplifiers have found widespread use in not only long-distance point-to-point optical fiber links but also in multiple-access networks to compensate for signal-splitting losses. The features of optical amplifiers has led to many diverse applications, each having different design challenges. Figure 6 shows general applications of optical amplifiers where the parameter G denotes gain.

In a single-mode link the effects of fiber dispersion may be small so that the main limitation to repeater spacing is fiber attenuation. Since such a link does not necessarily require a complete regeneration of the signal, simple amplification of the optical signal is sufficient. Thus an *inline optical amplifier* can be used to compensate for transmission loss and increase the distance between regenerative repeaters, as illustrated in Fig. 6a.

Figure 6b shows an optical amplifier being used as a front-end *preamplifier* for an optical receiver. In this way, a weak optical signal is amplified before photodetection so that the signal-to-noise ratio degradation caused by thermal noise in the receiver electronics can be suppressed. Compared with other front-end devices such as avalanche photodiodes or optical heterodyne detectors, an optical preamplifier provides a larger gain factor and a broader bandwidth.

Power or *booster amplifier* applications include placing the device immediately after an optical transmitter to boost the transmitted power, as Fig. 6c shows. This serves to increase the transmission distance by 10–100 km depending on the amplifier gain and fiber loss. As an example, using this boosting technique together with an optical preamplifier at the receiving end can enable repeaterless undersea transmission distances of 200–250 km. One can also employ an optical amplifier in a local-area network as a booster amplifier to compensate for coupler insertion loss and power-splitting loss. Figure 6d shows an example for boosting the optical signal in front of a star coupler.

2. LINK PERFORMANCE CHARACTERISTICS

The transmission characteristics are a major factor in determining what a signal looks like after it has propagated a certain distance. The three fundamental signal-distorting factors are attenuation, dispersion, and nonlinear effects in an optical fiber.

2.1. Attenuation

Attenuation of a light signal as it propagates along a fiber is an important consideration in the design of an optical communication system, since it plays a major role in determining the maximum transmission distance between a transmitter and a receiver. The basic attenuation mechanisms in a fiber are absorption, scattering, and radiative losses of the optical energy. Absorption is related to the fiber material, whereas scattering is associated both with the fiber material and with structural imperfections in the optical waveguide. Attenuation owing to radiative effects originates from perturbations (both microscopic and macroscopic) of the fiber geometry. As light travels along a fiber, its power decreases exponentially with distance. If $P(0)$ is the optical power in a fiber at the origin (at $z = 0$), then the power $P(z)$ at a distance z further down the fiber is

$$P(z) = P(0)e^{-\alpha_p z} \tag{1}$$

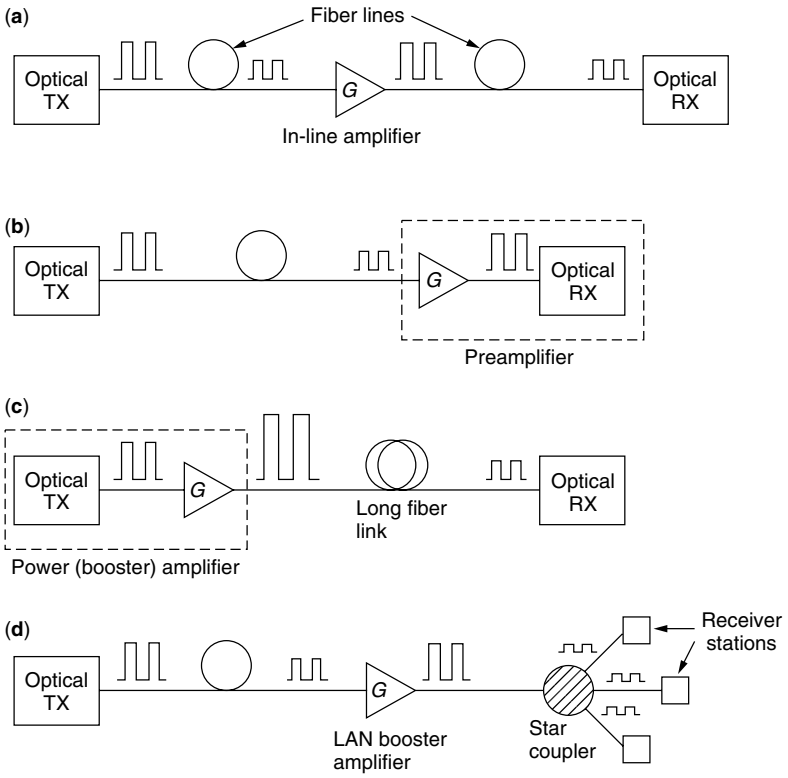


Figure 6. General applications of optical amplifiers.

where

$$\alpha_p = \frac{1}{z} \ln \left[\frac{P(0)}{P(z)} \right] \tag{2}$$

is the fiber *attenuation coefficient* given in units of, for example, reciprocal kilometers (km^{-1}). For simplicity in calculating optical signal attenuation in a fiber, the common procedure is to express the attenuation coefficient in units of *decibels per kilometer*, (dB/km). Designating this parameter by α, we have

$$\alpha(\text{dB/km}) = \frac{10}{z} \log \left[\frac{P(0)}{P(z)} \right] = 4.343\, \alpha_p(\text{km}^{-1}) \tag{3}$$

This parameter is generally referred to as the *fiber loss* or the *fiber attenuation*. Figure 3 illustrates the attenuation of optical power in a fiber as a function of wavelength.

2.2. Dispersion

Several different dispersion mechanisms in an optical fiber cause a light signal to become increasingly distorted as it travels along a fiber. These include material, waveguide, and polarization-mode dispersions [4,14]. *Material dispersion* arises because each optical pulse in a digital signal contains a small range of wavelengths. Since the refractive index n of silica glass varies slightly as a function of wavelength, the fundamental relationship for the velocity v in a material $v = c/n$, where c is the speed of light, shows that different parts of the pulse will travel at different speeds. Consequently, as a result of this material dispersion effect, a pulse will broaden as it travels along a fiber.

Waveguide dispersion occurs because a single-mode fiber confines only about 80% of the optical power to the core. The other 20% propagates in the cladding that surrounds the core. Dispersion arises since the light in the cladding sees a lower refractive index than in the core and thus travels faster than the light confined in the core. The amount of waveguide dispersion depends on the fiber design. Thus, through ingenious fiber construction, waveguide dispersion can be tailored to counteract the effects of material dispersion. In standard fiber designs, material and waveguide dispersions cancel at 1310 nm. To achieve zero total dispersion at 1550 nm, where the attenuation of a silica fiber is at its lowest point, the *dispersion-shifted fiber* was developed in the mid-1980s. This works well for single-wavelength operation, but is not desirable in WDM systems. Here nonlinear effects require different approaches, one of which is the dispersion management scheme described below.

Polarization-mode dispersion arises from the effects of fiber birefringence on the polarization states of an optical pulse [15]. *Birefringence* refers to slight variations in the indices of refraction along different axes of the fiber. This is particularly critical for high-rate, long-haul transmission links (e.g., 10 Gbps over tens of kilometers) that are designed to operate near the zero-dispersion wavelength of the fiber. Birefringence can result from intrinsic factors such as geometric irregularities of the fiber core or internal stresses on it. Deviations of less than 1% in the circularity of the core can already have a noticeable effect in a high-speed lightwave system. In addition, external factors such as bending, twisting, or pinching of the fiber can also lead to birefringence. Since all these mechanisms exist to some extent in any field-installed fiber, there will be a varying birefringence along its length.

A fundamental property of an optical signal is its polarization state. *Polarization* refers to the electric field orientation of a light signal, which can vary significantly along the length of a fiber. As shown in Fig. 7, signal energy at a given wavelength occupies two orthogonal polarization modes. A varying birefringence along the length of the fibre will cause all the polarization modes to travel at slightly different velocities and the polarization orientation will rotate with distance. The resulting difference $\Delta\tau$ in propagation times between the two orthogonal polarization modes will result in pulse spreading. This is known as *polarization-mode dispersion* (PMD).

2.3. Nonlinear Effects

Two different categories of optical nonlinear effects also have an adverse effect on signal quality [16–18]. The first category encompasses nonlinear inelastic scattering processes, which are interactions between optical signals and molecular or acoustic vibrations in a fiber. These

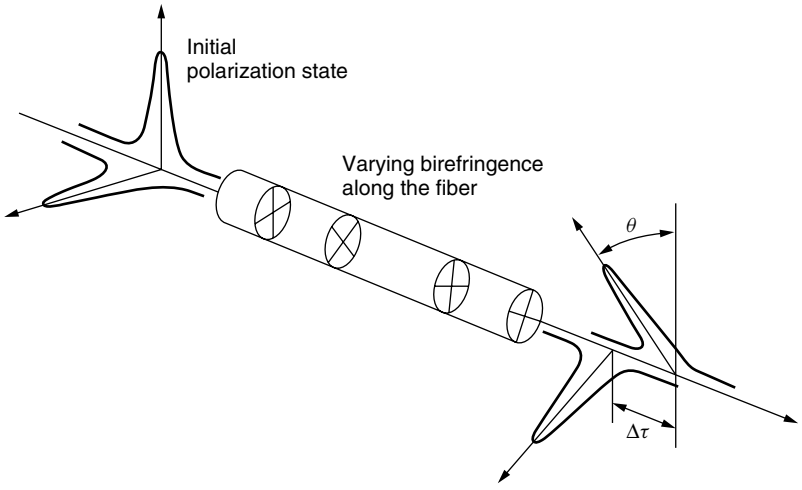


Figure 7. Signal energy at a given wavelength occupies two orthogonal polarization modes.

are stimulated Raman scattering (SRS) and stimulated Brillouin scattering (SBS). The second category involves nonlinear variations of the refractive index in a silica fiber that occur because the refractive index is dependent on intensity changes in the signal. This produces effects such as self-phase modulation (SPM), cross-phase modulation (XPM), and four-wave mixing (FWM). In the literature, FWM is also referred to as *four-photon mixing* (FPM), and XPM is sometimes designated by CPM.

SRS, SBS, and FWM result in gains or losses in a wavelength channel that are dependent on the optical signal intensity. These nonlinear processes provide gains to some channels while depleting power from others, thereby producing crosstalk between the channels. SPM and XPM affect only the phase of signals, which causes chirping in digital pulses. This can worsen pulse broadening due to dispersion, particularly in very high-rate systems (>10 Gbps). When any of these nonlinear effects contribute to signal impairment, an additional amount of power will be needed at the receiver to maintain the same BER (bit error rate) as in their absence. This additional power (in decibels) is known as the *power penalty* for that effect.

Stimulated Raman scattering is an interaction between lightwaves and the vibrational modes of silica molecules. If a photon with energy $h\nu_1$ is incident on a molecule having a vibrational frequency ν_m, the molecule can absorb some energy from the photon. In this interaction the photon is scattered, thereby attaining a lower frequency ν_2 and a corresponding lower energy $h\nu_2$. The modified photon is called a *Stokes photon*. Because the optical signal wave that is injected into a fiber is the source of the interacting photons, it is called the *pump wave*, since it supplies power for the generated wave. This process generates scattered light at a wavelength longer than that of the incident light. If another signal is present at this longer wavelength, the SRS light will amplify it and the pump wavelength signal will decrease in power. Consequently, SRS can severely limit the performance of a multichannel optical communication system by transferring energy from short-wavelength channels to neighboring higher-wavelength channels. This is a broadband effect that can occur in both directions. Powers in channels separated by up to 16 THz (125 nm) can be coupled through the SRS effect, thereby producing cross talk between wavelength channels. SRS may be controlled using fiber dispersion management techniques, as described below.

Stimulated Brillouin scattering arises when lightwaves scatter from acoustic waves. The resultant scattered wave propagates principally in the backward direction in single-mode fibers. This backscattered light experiences gain from the forward-propagating signals, which leads to depletion of the signal power. In silica this interaction occurs over a very narrow *Brillouin linewidth* of 20 MHz at 1550 nm. This means that the SBS effect is confined within a single wavelength channel in a WDM system, and thus accumulates individually for each channel. System impairment starts when the amplitude of the scattered wave is comparable to the signal power. For typical fibers the threshold power for this process is around 10 mW for single-fiber spans. In a long fiber chain containing optical amplifiers, there are normally optical isolators to prevent backscattered signals from entering the amplifier. Consequently, the impairment due to SBS is limited to the degradation occurring in a single amplifier-to-amplifier span. Several schemes are available for suppressing the effects of SBS [4].

The refractive index n of the glass material in an optical fiber has a weak dependence on optical intensity (equal to the optical power per effective area in the fiber). Since the intensity varies at the leading and trailing edges of each optical pulse in a digital datastream, the nonlinearity in the refractive index produces a carrier-induced phase modulation of the propagating signal. Consequently, parts of the pulse undergo a frequency shift in a process called *self-phase modulation* (SPM). As a result of dispersion in the fiber, this shift is transformed into pulse distortion. The effects of SPM can be reduced by maintaining a low overall dispersion in a fiber link.

In WDM systems, the refractive index nonlinearity gives rise to *cross-phase modulation* (XPM), which converts power fluctuations in a particular wavelength channel to phase fluctuations in other copropagating channels. This can be greatly mitigated in WDM systems operating over standard nondispersion-shifted single-mode fiber, but can be a significant problem in WDM links operating at 10 Gbps and higher over dispersion-shifted fiber. To mitigate XPM effects, the dispersion should be high, since pulses in different channels travel at different speeds, so that they walk through each other quickly, thereby minimizing their interactions.

Four-wave mixing (FWM) is a third-order nonlinearity in silica fibers, which is analogous to intermodulation distortion in electrical systems. The FWM effect arises from the beating between two or more channels, which creates new tones at other frequencies. When these new frequencies fall in the transmission window of the original frequencies, it can cause severe crosstalk. For DWDM systems, FWM can cause the highest power penalty of all the nonlinear effects. The efficiency of four-wave mixing depends on fiber dispersion and the channel spacings. Since the dispersion varies with wavelength, the signal waves and the generated waves have different group velocities. This destroys the phase matching of the interacting waves and lowers the efficiency at which power is transferred to newly generated frequencies. The higher the group velocity mismatches and the wider the channel spacing, the lower the effects of four-wave mixing.

2.4. Dispersion Management

Using current fiber designs, high-speed WDM systems are limited by nonlinear effects and dispersion. To mitigate these effects, dispersion management techniques are being used to maintain moderate dispersion locally and near-zero dispersion globally across the entire link [19–21]. This needs to be implemented across the wide spectrum of wavelengths used in WDM systems. To achieve this, one may use passive *dispersion compensation*. This consists of inserting into the link a loop of fiber having a dispersion characteristic that negates the accumulated dispersion of the transmission fiber. The fiber loop is referred to as a *dispersion-compensating fiber* (DCF). If the transmission

fiber has a low positive dispersion [say, 2.3 ps/(nm · km)], then the DCF will have a large negative dispersion [say, −16 ps/(nm · km)].

With this technique, the total accumulated dispersion is zero after some distance, but the absolute dispersion per length is nonzero at all points along the fiber. The nonzero absolute value causes a phase mismatch between wavelength channels, thereby destroying the possibility of effective FWM production.

3. MEASUREMENT METHODOLOGIES

The design and installation of an optical fiber communication system require measurement techniques for verifying the operational characteristics of the constituent components [4,22]. In addition to optical fiber parameters, system engineers are interested in knowing the characteristics of passive splitters, connectors, and couplers, and electrooptic components, such as sources, photodetectors, and optical amplifiers. Furthermore, when a link is being installed and tested, the operational parameters of interest include bit error rate, timing jitter, and signal-to-noise ratio as indicated by the eye pattern. During actual operation, measurements are needed for maintenance and monitoring functions to determine factors such as fault locations in fibers and the status of remotely located optical amplifiers.

4. FURTHER INFORMATION

Many of the concepts covered in this article are described in more detail elsewhere in this encyclopedia. For example, see: articles on nonlinear effects in fibers, optical amplifiers, optical couplers, optical fiber dispersion, optical filters, optical networks, optical receivers, optical transmitters, standards, and wavelength-division multiplexing.

BIOGRAPHY

Gerd Keiser is the founder and president of PhotonicsComm Solutions, Inc., Newton Center, Massachusetts, a firm specializing in consulting and education for the optical communications industry. He has 25 years experience at Honeywell, GTE, and General Dynamics in designing and analyzing telecommunication components, links, and networks. He is the author of the books *Optical Fiber Communications* (3rd ed. 2000) and *Local Area Networks* (2nd ed. 2002) published by McGraw-Hill. Dr. Keiser is an IEEE fellow and received GTE's prestigious Leslie Warner Award for work in ATM switch development. He earned his B.A. and M.S. degrees in mathematics and physics from the University of Wisconsin and a Ph.D. in solid state physics from Northeastern University, Boston, Massachusetts.

BIBLIOGRAPHY

1. D. J. H. Maclean, *Optical Line Systems*, Wiley, Chichester, UK, 1996 (this book gives a detailed discussion of the evolution of optical fiber links and networks).
2. R. Ramaswami and K. N. Sivarajan, *Optical Networks*, 2nd ed., Morgan Kaufmann, San Francisco, 2002.
3. J. Hecht, *City of Light: The Story of Fiber Optics*, Oxford Univ. Press, 1999.
4. G. Keiser, *Optical Fiber Communications*, 3rd ed., McGraw-Hill, Burr Ridge, IL, 2000.
5. K. C. Kao and G. A. Hockman, Dielectric-fiber surface waveguides for optical frequencies, *Proc. IEE* **133**: 1151–1158 (July 1966).
6. F. P. Kapron, D. B. Keck, and R. D. Maurer, Radiation losses in glass optical waveguides, *Appl. Phys. Lett.* **17**: 423–425 (Nov. 1970).
7. S. Tsuda and V. L. da Silva, Transmission of 80 × 10 Gbps WDM channels with 50-GHz spacing over 500 km of LEAF fiber, *Tech. Digest IEEE/OSA Optical Fiber Commun. Conf.*, March 2000, pp. 149–151.
8. R. C. Alferness, H. Kogelnik, and T. H. Wood, The evolution of optical systems: Optics everywhere, *Bell Labs Tech. J.* **5**: 188–202 (Jan.–March. 2000).
9. Special issue on "Undersea Communications Technology," *AT&T Tech. J.* **74**: (Jan./Feb. 1995).
10. G. E. Keiser, A review of WDM technology and applications, *Opt. Fiber Technol.* **5**: 3–39 (Jan. 1999).
11. S. V. Kartalopoulos, *Introduction to DWDM Technology*, IEEE Press, New York, 2000.
12. E. Desurvire, *Erbium-Doped Fiber Amplifiers*, Wiley, New York, 1994.
13. H. Masuda, Review of wideband hybrid amplifiers, *Tech. Digest IEEE/OSA Optical Fiber Commun. Conf.*, March 2000, pp. 2–4.
14. P. Hernday, Dispersion measurements, in D. Derickson, ed., *Fiber Optic Test and Measurement*, Prentice-Hall, Upper Saddle River, NJ, 1998.
15. C. D. Poole and J. Nagel, Polarization effects in lightwave systems, in I. P. Kaminow and T. L. Koch, eds., *Optical Fiber Telecommunications — III*, Vol. A, Academic Press, New York, 1997, Chap. 6, pp. 114–161.
16. G. P. Agrawal, *Nonlinear Fiber Optics*, 2nd ed., Academic Press, New York, 1995.
17. F. Forghieri, R. W. Tkach, and A. R. Chraplyvy, Fiber nonlinearities and their impact on transmission systems, in I. P. Kaminow and T. L. Koch, eds., *Optical Fiber Telecommunications — III*, Vol. A, Academic Press, New York, 1997, Chap. 8, pp. 196–264.
18. E. Iannone, F. Matera, A. Mecozzi, and M. Settembre, *Nonlinear Optical Communication Networks*, Wiley, New York, 1998.
19. B. Jopson and A. H. Gnauck, Dispersion compensation for optical fiber systems, *IEEE Commun. Mag.* **33**: 96–102 (June 1995).
20. L. Grüner-Nielsen et al., Dispersion compensating fibers, *Opt. Fiber Technol.* **6**: 164–180 (April 2000).
21. M. Murakami, T. Matsuda, H. Maeda, and T. Imai, Long-haul WDM transmission using higher-order fiber dispersion management, *J. Lightwave Technol.* **18**: 1197–1204 (Sept. 2000).
22. D. Derickson, ed., *Fiber Optic Test and Measurement*, Prentice-Hall, Upper Saddle River, NJ, 1998.

OPTICAL FIBER LOCAL AREA NETWORKS

MEHDI SHADARAM
VIRGILIO E. GONZALEZ-LOZANO
University of Texas at El Paso
El Paso, Texas

1. INTRODUCTION

Technological advances in the analog, digital, and photonic systems have transformed communications into a highly dynamic field. Nowadays, communication between computers is an essential part of modern living, and telecommunications is one of world's fastest-growing industries. Most advances in this field are driven by social, economical, political, and technological reasons. It is a well-known fact that without a reliable communication infrastructure, nations cannot retain a prosperous economy. The need for computer and communication engineers to implement new ideas and to satisfy the growing demand for higher bandwidth is apparent more than ever. The data rate of computer networks used in university campuses, hospitals, banks, and elsewhere is doubling almost every year. The data rate has reached a point where a single transmission medium such as twisted-pair or coaxial cable is not capable of transmitting the load. In order to avoid multiple-cables laying beneath the ground and overloading buildings with wires, the need for one single transmission medium that can convey up to several gigabits of information per second is necessary. Since the early 1970s optical fibers have been utilized as transmission media in long- and short-distance communication links. Typical optical fiber attenuation has decreased from several dB/km at 0.8 μm wavelength in the early seventies to about 0.1 dB/km at 1.55 μm wavelength as we enter the new millennium. During the same period, the capacity of optical fibers has increased from several Mbps/km (megabits per second per kilometer) to $\sim$300 Gbps/km. Because of gradual maturing of multiwavelength optical fiber systems, wavelength-division multiplexing (WDM), and development of zero-dispersion optical fibers (solitons), the capacity of a single optical fiber could reach thousands of Gbps in the near future. Small size, light weight, and immunity to electromagnetic interference noise also provide a crucial advantage for optical fibers over other media. Although cost of implementing fiber-based systems still exceeds the cost to deploy other systems such as coaxial or twisted pair, prices for optical fiber systems are dropping as rapidly as 30% per year.

2. WHAT IS AN OPTICAL FIBER LOCAL AREA NETWORK?

Organizations such as universities, hospitals, banks, or even small offices use computers for a variety of applications. Very often different users within these establishments need to share data. Thus, for a reliable and high-speed data transfer, computers are connected through a network of point-to-point communication links. These types of networks are usually referred to as *local-area networks* (LANs). The size of a LAN can vary anywhere from two computers connected to one another in a room to several thousand computers connected together in a large campus. The distance between nodes within a LAN can vary from a few meters to as much as 2 km. If computers are connected via optical fiber links, the LAN is referred to as *optical fiber LAN*. Optical fiber cables conduct light along a thin solid cylinder-shaped glass or plastic fiber, known as a *core*. The core is surrounded by *cladding*, which in turn is surrounded by a plastic sheath as shown in Fig. 1a. Depending on the fiber type, the core diameter can be in the range of 8–62 μm. The typical diameter of cladding for optical fibers used in telecommunication is about 125 μm. In order to confine the light within the core, as shown in Fig. 1b, the refractive index of the core is slightly higher than the refractive index of the cladding, as can be seen in Fig. 1c. The block diagram shown in Fig. 2 exhibits a typical optical fiber link used between two nodes within the network. The link typically consists of a laser diode (LD) or light-emitting diode (LED) in the transmitter unit, which launches an optically modulated signal into an optical fiber cable. At the receiver site, the cable is terminated by a photodiode, which converts the optical signal into an electric current.

In the early days of optical fiber technology, because of major developments in gallium arsenide devices, most fiber links were operating at short wavelengths ($\sim$0.8 μm). These links could carry bit rates of $\leq$100 Mbps with an attenuation of $\sim$10 dB/km. Nowadays, most optical fiber links operate at around 1.3 μm, minimum dispersion wavelength; or 1.55 μm, minimum attenuation wavelength. Development of dispersion-shifted fibers has made it possible for fibers to exhibit minimum dispersion and minimum attenuation at 1.55 μm. Links employing dispersion-shifted fibers are commonly used for long-distance communication purposes.

The main purpose of a LAN is to share resources such as files, printers, programs, and databases among the nodes in the network. The initial LAN designs were intended for communication between computers at short distances. However, the length of the links has grown from a few meters to a few kilometers.

To reach longer distances there are wide-area networks (WANs), which typically employ the infrastructure of a telecom service provider or carrier. A single network manager usually controls the LAN operation, but the WAN requires the coordination between the service provider and the LAN administrator. Very often, carrier networks require an initial setup of a connection between the end nodes, while LANs do not require that setup.

Through the years the capacity and the transmission distance of the networks have increased. The process has created a hybrid type of network between a LAN and a WAN. This type of network is called a *metropolitan-area network* (MAN). The MAN is capable of covering an entire city using the simplicity of the LAN protocols at 100 Mbps and above. On the other hand, there is a new specialized small network called *storage-area network* (SAN). The purpose of a SAN is to allow very-high-speed communication between computer processors and dedicated peripherals such as large disk arrays.

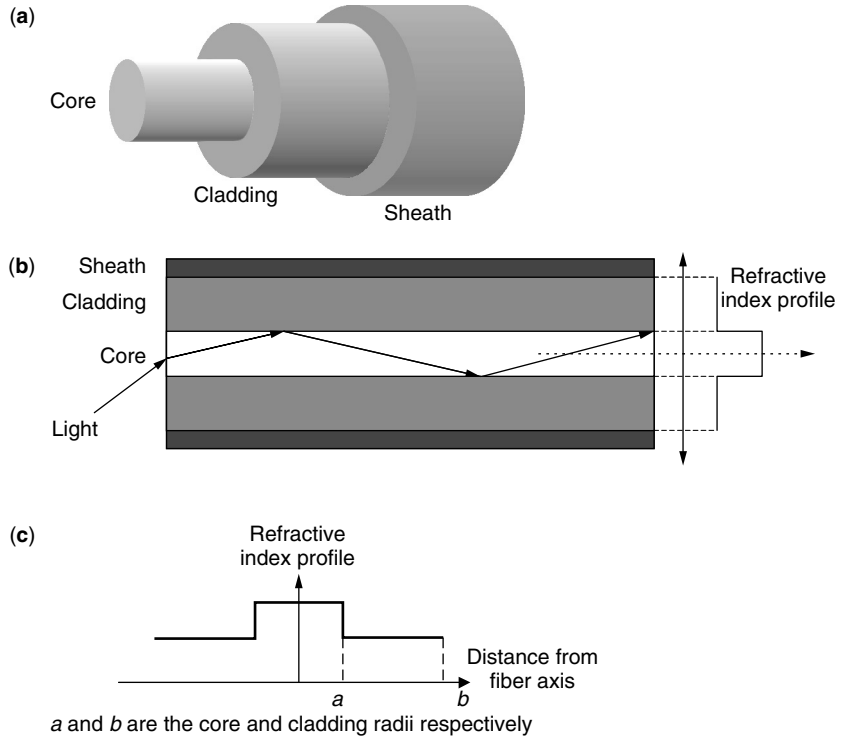


Figure 1. Optical fiber.

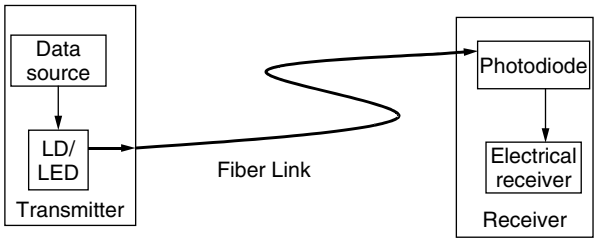


Figure 2. Typical optical fiber link.

3. NETWORK TOPOLOGIES

As pointed out previously, a network allows users to share information in an efficient, fast, and reliable manner. As images, audio, and video files become a regular part of the datastream, the need for networks with higher bandwidth becomes more apparent. Increasing the number of point-to-point links within a network or increasing the bandwidth of point-to-point links can increase the network bandwidth. Optical fiber is usually favored over other cables, since it is capable of very high data transfer rates. An optical link can replace an electric link as long as all electrical interface requirements such as voltage or current signal levels, timing, and control signals are met.

Implementation of an optical fiber LAN may follow one of two approaches. The first method involves the creation of a completely new network with optical fiber links. The second approach requires the replacement, within a conventional LAN, of electric links with optical links while meeting all original interface requirements. As a result, many LANs expand their reach by adding special repeaters with fiber segments.

There are two types of signal couplings to the fiber, known as active and passive. Figure 3a shows that, for active coupling, the network behaves like a series of point-to-point links and each node must be operational to maintain the network working. For passive coupling, stations broadcast the signals into a common fiber and each node captures only a small amount of their power. As seen in Fig. 3b, an inactive station does not affect the operation of the network. An alternative for inactive nodes is to include an optical bypass at each station, as shown in Fig. 4.

The network topology establishes a path for the flow of information between the transmitter and the rest of the network. The ring, star, and bus network topologies have become popular for most optical fiber LANs. For example, rings are used in the Fiber Distributed Data Interface (FDDI), the IEEE 802.6 Dual Queue Dual Bus (DQDB) standard uses a dual bus, and star configurations are encountered in switched systems. The LAN logical topology may differ from the physical cable layout. The

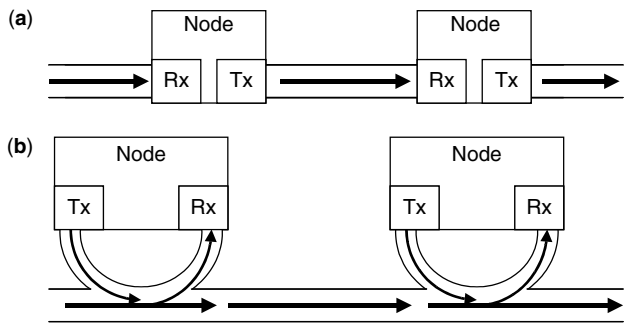


Figure 3. Types of coupling for optical links: (**a**) active and (**b**) passive coupling.

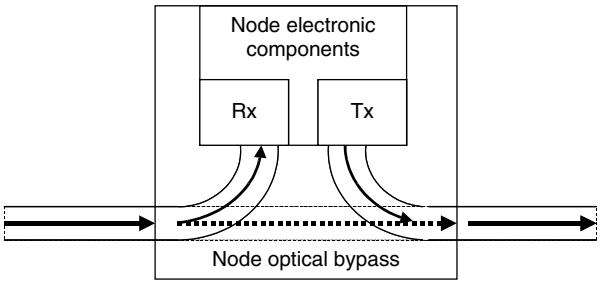


Figure 4. Optical bypass of a node.

budget, time, and resources usually dictate how the cable should be laid. For example, a ring topology can be folded to resemble two buses going in opposite directions using different fiber strands.

The technique to control the fair share of the transmission medium is defined by the medium access control (MAC) protocol and is dependant on the topology. The most common types of MAC protocols use token-passing, reservation, or collision avoidance techniques. The token-passing technique allows only the station holding the token to transmit at a time. The token is circulated through all the stations with a predefined mechanism that prevents unfair behavior and restores lost tokens. The reservation mechanism requires an arbiter that collects reservation requests from the stations and then allocates turns to transmit; this is commonly used in networks with asymmetric traffic needs. The collision mechanism allows stations to talk at any time, but two nodes trying to transmit simultaneously produce a collision. In that case, they cease transmission and wait a random time before retransmission. This method causes one of them to start before the other, reducing the probability of a new collision.

3.1. Ring Topology

Logic ring architecture, as can be seen in Fig. 5, establishes the circulation of information in a specific direction around the ring. The ring passes through all the stations in the LAN, enabling all of them to receive the same message until it completes the loop. The vulnerability of this network is that a single interruption in the ring disrupts the whole LAN. To maintain the reliability, the most common approaches are to create a second counterrotating ring for restoration, as shown in Fig. 6, or to add a bypass device for the damaged segment.

The most common mechanism to control access to the medium is the use of token-passing techniques. The ring nature simplifies the process to circulate the token among the stations. Only the station that possesses the token may transmit, and the rest remain silent. When the turn finishes, the station passes the token to the next on the ring, giving an opportunity to all nodes to transmit.

3.2. Bus Topologies

The bus topology establishes a sequence of nodes in line as exhibited in Fig. 7. The more commonly used access control mechanisms are collision detection and reservation of resources. Stations in a bus could use either active or passive coupling. The transmission into the bus

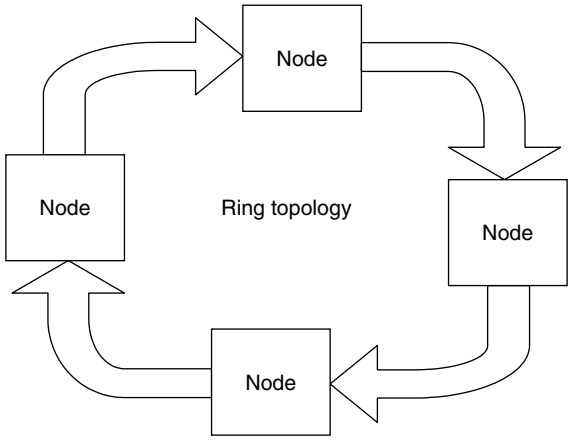


Figure 5. Ring topology.

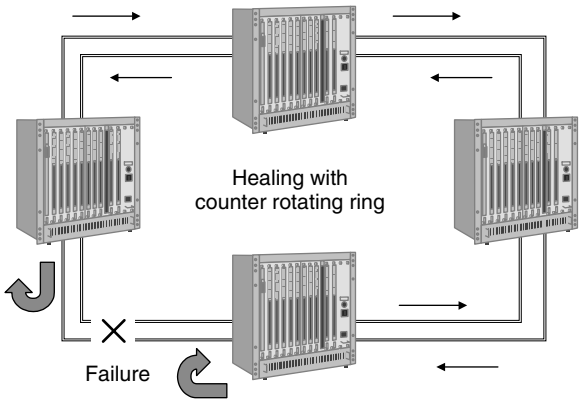


Figure 6. Self-healing ring.

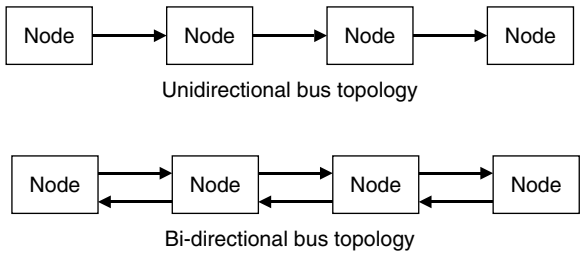


Figure 7. Bus topologies.

can be made in both directions; however, it is usually unidirectional for fiberoptic networks. For unidirectional fiber segments we need a return path, so typically there are two buses going in opposite directions. A station in a logic bus transmits information into the cable, and the remaining stations receive the transmission downward. The network will be divided into two independent segments if there is a break in the bus, unless there is an alternate path to rejoin the segments together.

3.3. Star Topologies

The star topology links several stations to a single central point. The connectivity in the network could be active

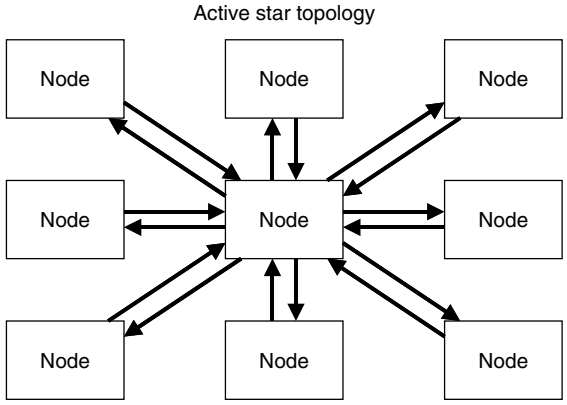


Figure 8. Active star topology.

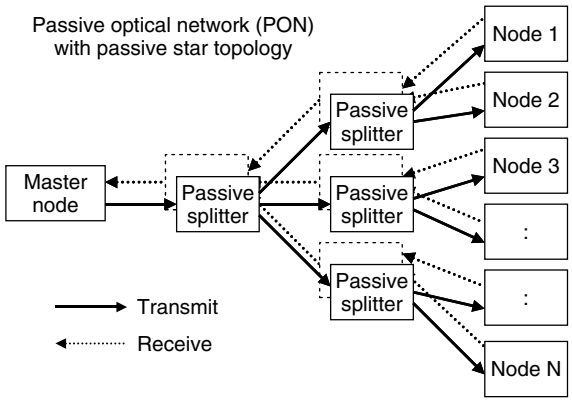


Figure 9. Passive star topology.

or passive, as illustrated in Figs. 8 and 9, respectively. Most configurations require the central station to be the controller for the rest of the network. It acts as a switch for messages between stations or performs as an arbiter for the allocation of resources. Fewer star networks distribute the MAC functionality to the nodes using token-passing or collision detection techniques. The more vulnerable point of the network is the central station because a break affects all the stations below the point of rupture.

The use of passive optical components, splitting the light among different fibers, allows several stations to receive the same broadcast from a single transmitter as shown in Fig. 10. These types of networks are commonly known as *passive optical networks* (PONS). The PON architecture normally requires a central station that controls the rest of the network. The access control is made by reservation of a resource, either a time slot or a whole optical channel.

Ethernet-type networks have collapsed the bus to a single-hub device, similar to the twisted-pair cabling option of Ethernet (known as *10BASE-T*). In addition, they have extended the individual links for each station using optical fiber. Therefore, the physical layout is a star and the links are point-to-point.

4. NETWORK DESIGN CONSIDERATIONS

Two important parameters need to be evaluated before designing a network: power loss and dynamic range. The assessment of these constraints depends on the network topology. In the bus topology, the optical signal is typically tapped by using an optical coupler at each node. In the ring topology, the coupling between the ring and the node can be either passive or active as illustrated in Fig. 3. In the star topology with N nodes, an $N \times N$ optical fiber star coupler is usually used to distribute the signal from one input to N outputs equally.

If we assume that couplers used in a bus topology couple C percent of the power from the bus to a node with a typical coupling loss of α dB, the total power coupling from the bus to a node will be $C \times 10^{-\alpha/10}$. In this case, the maximum power loss from one node to another node will be

$$L_{\text{bus}} = \frac{10^{\alpha N/10}}{(1-C)^{N-2}C^2}$$

$$L_{\text{bus,dB}} = 10(2-N)\log_{10}(1-C) - 20\log_{10} C + \alpha N \text{ dB} \quad (1)$$

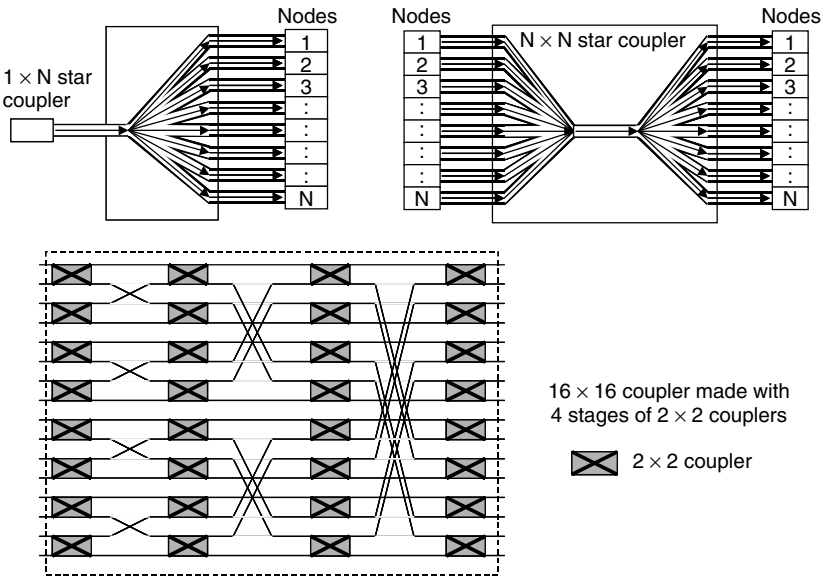


Figure 10. Optical star couplers.

where N is the total number of nodes. The maximum power loss occurs when the transmitting and the receiving nodes are at two opposite ends of the bus. As can be seen, the maximum power loss is a function of coupling ratio C and N. The optimum value of C, which minimizes the maximum power loss, is given by

$$C_{\text{opt}} = \frac{2}{N} \tag{2}$$

Substituting C_{opt} from Eq. (2) into Eq. (1) results in

$$L_{\text{bus,dB,min}} = 20\log_{10}(N-2) - 10N\log_{10}\left(\frac{N-2}{N}\right) + \alpha N - 6 \text{ dB} \tag{3}$$

The dynamic range (DR) is defined as the ratio of the maximum received power to the minimum received power by a node. In the single-bus topology, the maximum received power occurs when the signal comes from the adjacent node, and the minimum received power occurs when the receiving and transmitting nodes are farthest apart from each other. The DR for bus topology can be evaluated using the following procedure:

$$P_{\max} = P_0 C^2 10^{-2\alpha/10}$$
$$P_{\min} = P_0 C^2 (1-C)^{N-2} 10^{-N\alpha/10}$$
$$DR_{\text{bus}} = \frac{P_{\max}}{P_{\min}} = (1-C)^{2-N} 10^{(\alpha/10)(N-2)}$$
$$DR_{\text{bus,dB}} = \alpha(N-2) - 10N\log(1-C) + 20 \text{ dB}$$

where P_0 is the transmitted power.

An $N \times N$ star coupler typically consists of $\log_2 N$ stages of 2×2 couplers as shown in Fig. 10. In the star topology, the transmitted signal is equally distributed to N nodes. Thus, the power loss from a transmitter at one input to a receiver at one of the outputs can be evaluated using the following equation:

$$L_{\text{star}} = N10^{[(\alpha \log_2 N)/10]}$$
$$L_{\text{star,dB}} = (3+\alpha)\log_2 N \text{ dB} \tag{4}$$

where α is the coupling loss at each 2×2 stage. Since the transmitted power is equally divided among the receiving nodes, the DR in star topology is unity.

5. NETWORK PROTOCOLS

Most of the current LAN protocols are derived from the standards defined in the IEEE 802 series [1]. They specify the packet formats and medium access techniques for various types of transmission media. However, there are other types of networks defined by Industry associations and standards bodies. Table 1 shows the main characteristics of popular optical fiber LANs and we will describe the more relevant types below.

5.1. FDDI

In the late 1980s, The American National Standard Institute (ANSI) defined the protocol denominated Fiber Distributed Data Interface (FDDI) [2]. The accredited standards committee (ASC X3T9.5) had the objective to develop an inexpensive high-speed optical network functioning as the backbone of other slower LANs. Later,

Table 1. Characteristics of Common Optical LANs

Network Type	Fiber Type	Speed (Mbps)	Coding	Maximum Segment Distance	Topology	Access Method
FDDI	Multimode, 1300 nm	100	4B/5B NRZI	2 km (maximum 100 nodes)	Ring	Token passing
Ethernet 10BASE-T (fiber)	Multimode	10	Manchester	500 m	Active star	CSMA/CD
Ethernet 10BASE-FP	Multimode	10	Manchester	1 km (maximum 33 nodes)	Passive star	CSMA/CD
Ethernet 10BASE-FL/FB	Multimode or single-mode	10	Manchester	2 km	Active star	CSMA/CD
100BASE-FX/SX	Multimode	100	4B/5B NRZI	2 km	Active star	CSMA/CD
1000BASE-SX	Multimode	1000	8B/10B	275–550 m	Active star	CSMA/CD
1000BASE-LX	Multimode or single-mode	1000	8B/10B	550 m–5 km depending on fiber type	Active star	CSMA/CD
ATM LAN emulation	Single-mode	43, 155, 622	Any valid SONET, SDH, or DS3 line	Any valid SONET, SDH, or DS3 line	Active star	Switched central control
Fiber channel	Multimode or single-mode	100, 200, 400 & 800	8B/10B	175 m–10 km depending on fiber type	Active star	Switched central control

the International Standards Organization (ISO) published the same specifications under the ISO 9314 series.

FDDI specifies a 100-Mbps LAN using ring topology and token-passing access control. It employs multimode optical fiber and LEDs because they are more economical than laser diodes and single-mode fibers. The maximum distance between nodes is 2 km, and there is a maximum of 100 nodes per ring. The maximum frame size is 4500 bytes. The network is composed of two rings; the primary is for normal operation, and the secondary is reserved for restoration. Faults in the ring are resolved by sending the traffic through the secondary ring in the opposite direction, as shown in Fig. 6. Inactive nodes are optically bypassed, either internally or using a central hub with connection only to the primary ring.

5.2. IEEE 802.3 (Ethernet)-Type Protocols over Fiber

The Ethernet, developed by Xerox, is the basis for the IEEE 802.3 [1] standard using the CSMA/CD technique. It is the most popular protocol for LANs and has been improved to work at various speeds over many types of media. Since the first Ethernet implementations, there have been optical transceivers extending point-to-point connections [3]. However, there are newer standards to support higher speeds over several types of fiber. Most implementations use point-to-point fiber links in an active star configuration. The hubs convert the optical signals to electronic format and process them similar to other Ethernet LANs. The exception is a passive star defined in 10BASE-FP protocol. Lasers and single-mode fibers allow longer links for optical LANs; however, they are limited because they need to detect collisions. To overcome the problem, many networks employ bridges or switches in point-to-point links. Some equipment manufacturers have created proprietary solutions to reach distances of several hundreds of kilometers.

All the specifications employ the same frame format as 802.3. Only one station may transmit at a time, and if two nodes generate a collision, the hub sends a "collision presence" signal to the remaining stations. The standards 10BASE-FL and 10BASE-FB define 10-Mbps networks that may use multimode or single-mode fiber links. The difference between the standards is the retiming provided by the repeaters in 10BASE-FB. The 100-Mbps standards tried to reuse existing elements from other networks. For example, 100BASE-FX uses the same optical components similar to FDDI, and 100BASE-SX employs fiber compatible with the 10BASE-FL standard. The Gigabit Ethernet standard (1000BASE-SX/LX) specifies 1 Gbps transmission over multimode and single-mode fibers. There is an effort, by telecommunications service providers, to deploy 10-Gbps Ethernets. The standards are based on optical components similar to SONET OC-192 systems.

5.3. ATM Lan Emulation and Fiber Channel

5.3.1. ATM. The asynchronous transfer mode (ATM) protocol was designed by the telecom industry to handle all types of communication. It fragments the information into fixed-size packets, called cells, to work with real-time applications and bursty traffic. The definitions were made to support the backbone for carriers and the LANs in the corporations. ATM was designed to work principally on top of synchronous optical network (SONET) [4] links; the advantage is that the links can span very large distances using standard carrier equipment. ATM was expected to gradually replace LAN [5] protocols; however, the combination of Ethernet with TCP/IP dominated the LAN market.

ATM is a switching protocol that requires the establishment of virtual circuits between stations before a communication can be effected. The network administrators manually configure typical ATM implementations; however, LANs operate under a connectionless approach. The nodes in a LAN just send packets with enough information to reach the destination without the need to negotiate circuit establishment. ATM emulates a LAN relying in a signaling protocol that manages the connections.

The LAN emulation (LANE) standards define a mechanism in which the ATM stations behave like a LAN. The nodes employ a server that controls the establishment of the circuits and keeps track of the stations registered as members of the LAN. The individual nodes need to register their address with this server before requesting a connection to send packets. There is a second server used for broadcasting; it replicates a broadcast packet and sends it to all stations. This solution offers high-performance links between stations because there is no contention for resources in the switch. However, companies have limited its use because LAN switches are simpler and inexpensive compared to ATM switches.

5.3.2. Fiber Channel. The Fiber Channel standard was created to support communication between computers and intelligent distributed peripherals at short distances. When peripherals required higher transfer speeds, the interfaces using copper cable were more difficult to implement and the distances were reduced. The proposed solution for the problem defined a switched network with a star configuration. Each optical link is a point-to-point connection between switches and nodes. The standard was designed to encapsulate other types of protocols in point-to-point connections (ATM, SCSI, HPPI, IP, IEEE 802, etc). Fiber Channel supports multiple speeds, multimode or single mode fiber, and distances ranging from few meters up to 10 km. This protocol is normally used to communicate storage-area networks.

6. MULTIWAVELENGTH NETWORKS

WDM is a technique used to multiplex several optical channels through the same optical fiber. The operation is equivalent to combining and splitting light rays with different colors using a prism. A dense WDM system can be viewed as a parallel set of optical channels, each operating at a different wavelength, as illustrated in Fig. 11. This technology can increase the capacity of existing networks without the need for expensive additional cabling, reducing the cost of upgrades.

The bandwidth required for a channel transmission is in the order of gigahertz; however, the optical fiber

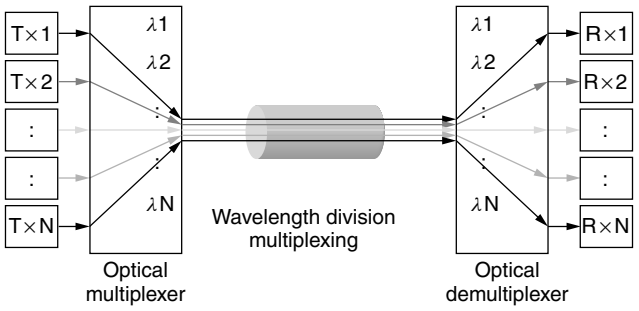


Figure 11. Wavelength-division multiplexing.

supports many terahertz in the available optical windows. WDM generally exploits the capacity of the optical fiber at 1300- and 1550-nm optical windows. The earlier WDM systems used only two channels, one at each window. The number of channels was increased later with the aid of optical filters, reaching separation between wavelengths in the range of tens of nanometers. More recently, the use of coherent techniques allowed separations of less than 1 nm. With this technology, the capacity of a single optical window expands to several dozens of channels known as *dense WDM* (DWDM).

There are several methods to allocate channels in WDM networks, where each wavelength corresponds to a channel. The major problem for these networks is to tune transmitters and receivers to different wavelengths.

Some methods propose fixed wavelengths assigned to each transmitter or each receiver. In order to establish communication, the other device will change its operating wavelength to match the desired station. This tuning process, in only one type of device, simplifies the network construction. However, it restricts the number of nodes to be the same as wavelengths available, and is subject to conflicts when two stations try to access the same destination.

In multihop [9] configurations the wavelength assignment is fixed for all the devices, and each station has several transmitters and receivers. A node can use other machines as repeaters, when it needs to talk to a station with an incompatible wavelength. This limits the number of nodes that can be reached from a particular source. However, the assignment of wavelengths is designed to create a sequence between stations forming a logical ring. This method simplifies the construction of the devices and reduces the conflict probability. On the other hand, it increases the traffic and the delay because it may require several retransmissions of the information.

Another technique enables all stations to tune the transmitters and receivers to any available wavelength. The method allows direct communication between all the stations and reduces the probability of blocking. Some disadvantages are complexity of the protocols required and more expensive devices. The nodes normally use a central control station to assign the wavelengths; however, they can employ techniques such as predefined sampling of channels, ALOHA protocol variants [6], and token passing. Those methods are not commonly used because they increase the delay and complexity of nodes.

In some networks, the signaling for call control is associated with the channel to use. This method blocks a resource just for the transmission of control information. A more efficient method for signaling employs a dedicated control channel that is independent of the information channels.

6.1. Phases for the Application of WDM Technology

The introduction of the WDM technology into optical fiber LANs, will follow an assimilation process similar to long-distance networks. At the beginning, it will serve the demand for increased bandwidth on some point-to-point links. Later, it will permeate to the backbone facilities, and finally will reach user interfaces. The evolution of DWDM technology seems likely to follow the stages shown below:

1. Increase the transmission capacity for congested point-to-point links.
2. Creation of an optical backbone that should add restoration capabilities.
3. Introduction of optical cross-connects allowing entire optical paths between end devices; however, the circuits are manually configured.
4. Usage of optical circuit switches. The stations will establish the optical circuits automatically, but will require sophisticated signaling protocols.
5. Introduction of optical packet routing. In the later stages the optical networks will be able to route individual packets without converting them to an electric form.

There is a trend, in the communications industry to create an all-optical network. The process will take several years to widely deploy commercial products. Meanwhile several experimental networks have been created to study their behavior. Some examples are Bellcore's Lambdanet [7], Lightning [8], Columbia's Telecomm Center TeraNet [9], IBM's RAINBOW [10,11]; All Optical Networking Consortium (AON) [12,13], Stanford's STARNET [14], and the European Advanced Communications Technologies and Services KEOPS [15] project.

6.2. DWDM System Components

Optical transmission systems may be divided into transmitters, receivers, amplifiers, and passive components. Each has to meet particular requirements to work in a WDM environment. The transmitters used for conventional optical fiber systems are light-emitting diodes or lasers. The link distance limits the available bandwidth, mainly due to a phenomenon called *dispersion*. It is a function of the type of fiber and transmitters used. Short-distance systems, in the order of few hundred meters, may use LEDs. For longer-distance transmission, the links require low dispersion; hence the use of lasers. All DWDM systems require a light source with a very narrow linewidth. This is accomplished with special types of lasers or with external filters.

The receiver must have good sensitivity and fast response for high-speed systems. The selection of a particular wavelength is obtained using optical filters.

There are several types of fixed filters manufactured for a specific optical channel; however, there are experiments to obtain a tunable receiver that is suitable for mass production.

For DWDM the most common optical amplifier is the erbium-doped fiber amplifier (EDFA) [16]. It amplifies the optical power of all the channels simultaneously, eliminating the need for several electronic devices. Mixing different wavelengths requires the use of optical couplers [16]. These devices combine different optical wavelengths at the inputs and distribute the mixed signal to all the outputs.

The International Telecommunications Union (ITU) proposed, through the Study Group 15 Work Project 4, a standard for wavelength assignment. It is based on a frequency reference of 193.1 THz, and a range of frequencies from 191.1 to 196.5 THz spaced 100 or 200 GHz each as shown in Table 2 (~0.8- or 1.6-nm spacing from 1568.77 to 1525.66 nm). It is ITU-T recommendation G.692 [17], "Optical interfaces for multichannel systems with optical amplifiers." This standard allows the creation of compatible devices from different vendors that operate on the same wavelengths.

7. CONCLUSION

Homes and corporations use LANs more and more each day, and the Ethernet family of protocols is the fastest-growing segment in LANs. Because of the high-bandwidth demand and wide range of applications, Ethernet is evolving into a hybrid network. The applications with lower requirements normally use conventional copper wiring; however, some of them have started to transform into wireless LANs. Users that need the higher data rates employ Gigabit Ethernet and 10-Gigabit Ethernet. This is the segment that needs more benefits from optical LANs. One lesson learned during this evolution is to reuse the existing technology, as much as possible, to make the products commercially viable. Fast Ethernet employs several elements from FDDI, and there is a trend for 10-Gigabit Ethernet to use the same optical technology as SONET OC-192 transport. Several optical LAN protocols are technically superior compared to Ethernet; however, the market has put them in disuse.

More recently the WDM technology has been utilized for the MANs. More likely in few years this technology will enter the LAN environment as the need for bandwidth increases. The major drivers for optical fiber networks are the multimedia applications and the communication between computers and peripherals in distributed computing environments.

The service providers have recognized that corporations need to link their LANs without the hassle of converting protocols. Therefore, they are currently offering LAN access ports to the customers. All the users share a high-capacity backbone segmented with virtual LANs (VLANs). This allows the corporations to reduce the costs to interconnect sites at high speeds; however, they are

Table 2. DWDM Grid Defined by Recommendation ITU-T G.692 with a Central Frequency of 193.10 THz and Separation of 100 GHz Between Channels

Number	Frequency THz	Wavelength nm	Number	Frequency THz	Wavelength nm
1	191.10	1568.77	29	193.90	1546.12
2	191.20	1567.95	30	194.00	1545.32
3	191.30	1567.13	31	194.10	1544.53
4	191.40	1566.31	32	194.20	1543.73
5	191.50	1565.50	33	194.30	1542.94
6	191.60	1564.68	34	194.40	1542.14
7	191.70	1563.86	35	194.50	1541.35
8	191.80	1563.05	36	194.60	1540.56
9	191.90	1562.23	37	194.70	1539.77
10	192.00	1561.42	38	194.80	1538.98
11	192.10	1560.61	39	194.90	1538.19
12	192.20	1559.79	40	195.00	1537.40
13	192.30	1558.98	41	195.10	1536.61
14	192.40	1558.17	42	195.20	1535.82
15	192.50	1557.36	43	195.30	1535.04
16	192.60	1556.55	44	195.40	1534.25
17	192.70	1555.75	45	195.50	1533.47
18	192.80	1554.94	46	195.60	1532.68
19	192.90	1554.13	47	195.70	1531.90
20	193.00	1553.33	48	195.80	1531.12
21	*193*.10	*1552*.52	49	195.90	1530.33
22	193.20	1551.72	50	196.00	1529.55
23	193.30	1550.92	51	196.10	1528.77
24	193.40	1550.12	52	196.20	1527.99
25	193.50	1549.31	53	196.30	1527.22
26	193.60	1548.51	54	196.40	1526.44
27	193.70	1547.72	55	196.50	1525.66
28	193.80	1546.92	56	193.80	1546.92

not exposed to the security risks involved in other public networks.

BIOGRAPHIES

Mehdi Shadaram received his B.S.E.E. degree from the University of Science and Technology in Tehran in 1976, his M.S. and Ph.D. degrees from the University of Oklahoma, both in electrical engineering, in 1980 and 1984, respectively. Currently, he is the Schellenger endowed professor and the chairman of the Department of Electrical and Computer Engineering at the University of Texas at El Paso. His research activities are focused in the field of optical fiber communications and photonic devices. During the last few years, he has investigated the performance of analog optical fiber links, WDM networks, and application of tapered single-mode optical fibers. NASA, Jet Propulsion Laboratory, National Science Foundation, Office of Naval Research, Department of Defense, Texas Instruments, Nortel Networks, and Lucent Technologies have funded his research projects. He has published more than 60 articles all in his area of research, most of them in refereed journals and conference proceedings. Dr. Shadaram is a registered professional engineer in the state of Texas. He is a senior member of IEEE, member of the International Society for Optical Engineering, Optical Society of America, and Eta Kappa Nu. He has received numerous awards for teaching and research excellence. He is cited in Marquis *Who's Who in America*.

Virgilio Gonzalez received his B.S. degree in Electrical Engineering in 1988 and M.S. degree in Industrial Engineering in 1991 from the Instituto Tecnologico y de Estudios Superiores de Monterrey (ITESM-CEM), Mexico. Later he obtained his Ph.D. degree in electrical engineering from the University of Texas at El Paso, in 1999, with the dissertation "Performance Analysis of a Fiber Optic Local Area Network Based in DWDM and ATM." From 1989 to 1993 he worked at the ITESM-CEM as telecommunications director developing the communications infrastructure for the different campus of ITESM University system. He worked from 1996 to 2001 in Alestra, the AT&T subsidiary Carrier in Mexico, as technology planning manager. His main responsibilities were the architecture design, technology development, and testing for all network functions in the carrier national network. In Mexico, he established the first Internet connection to Mexico City in 1989, set up the largest computer network in the country in 1991, and deployed the first DWDM network in 1998. Since 2001, he has been a professor at the University of Texas at El Paso. His areas of interest are high-speed optical communications, multiservice networks, and data communications protocols.

BIBLIOGRAPHY

1. IEEE Standards Assoc., *Welcome to Get IEEE 802™* (Oct. 31, 2001), Homepage (online): *http://standards.ieee.org/getieee802/* (Dec. 17, 2001).
2. W. Stallings, *Local and Metropolitan Area Networks*, 6th ed., Prentice-Hall, Upper Saddle River, NJ, 2000
3. International Engineering Consortium, *IEC Online Education — Optical Ethernet* (n.d), Homepage (online): *http://www.iec.org/online/tutorials/opt_ethernet/* (Dec. 17, 2001).
4. W Stallings, *ISDN and Broadband ISDN, with Frame Relay and ATM*, Prentice-Hall, Upper Saddle River, NJ, 1999
5. N. Kavak, Data communication in ATM networks, *IEEE Network* **9**(3): 28–37 (1995).
6. G. E. Keiser, *Local Area Networks*, McGraw-Hill, New York, 1989, pp. 205–214.
7. M. S. Goodman et al., The lambdanet multiwavelength network: Architecture, applications, and demonstrations, *IEEE J. Select. Areas Commun.* **8**(6): 995–1004 (Aug. 1990).
8. P. W. Dowd, K. Bogineni, K. A. Aly, and J. Perreault, Hierarchical scalable photonic architectures for high-performance processor interconnection, *IEEE Trans. Comput.* **42**(9): 1105–1120 (Sept. 1993).
9. R. Gidron and A. Temple, TeraNet: A multi-hop multichannel ATM lightwave network, *Conf. Records IEEE OFC 95*, 1995.
10. N. R. Dono et al., A wavelength division multiple access network for computer communication, *IEEE J. Select. Areas Commun.* **8**(6): 983–994 (Aug. 1990).
11. E. Hall et al., The Rainbow-II gigabit optical network, *IEEE J. select. Areas Commun.* **14**(5): 814–823 (June 1996).
12. S. B. Alexander et al., A precompetitive consortium on wideband all-optical networks, *J. Lightwave Technol.* **11**(5/6): 714–735 (May/June 1993).
13. Consortium (July 30, 1997), Homepage. (online): *AON All-Optical Networking*. *http://www.ll.mit.edu/aon/* (Dec. 17, 2001).
14. T. K. Chiang et al., Implementation of STARNET: A WDM computer communications network, *IEEE J. Select. Areas Commun.* **14**(5): 824–839 (June 1996).
15. M. Renaud, F. Masetti, C. Guillemot, and B. Bostica, Network and system concepts for optical packet switching, *IEEE Commun. Mag.* **35**(4): (April 1997).
16. J. Gowar, *Optical Communication Systems*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1993.
17. ITU-T Recommendation G.692, *Optical Interfaces for Multichannel Systems with Optical Amplifiers*, Geneva: International Telecommunication Union, 1998.

OPTICAL FILTERS

LEON POLADIAN
University of Sydney
Eveleigh, Australia

1. INTRODUCTION

An optical filter introduces a wavelength- or frequency-dependent change in the amplitude or phase of an optical signal passing through it. They are important components in modern optical communications networks that exploit the ability to simultaneously transmit information on more than one wavelength along a single optical fiber or

link [wavelength-division multiplexed (WDM) networks]. These WDM networks require components that can manipulate, combine, change, and reroute information based on the wavelength of light; optical filters perform many of these functions [1–4].

Filters can operate as either selective or corrective devices, or often combining both attributes. *Selective* devices extract or separate an optical signal into separate components based on wavelength or frequency and are used in optical demultiplexing, add/drop filters, optical switching, and also to create narrow wavelength selective mirrors in various lasers or other optical cavity–based devices. *Corrective* devices are used to adjust the amplitude or phase of the optical signal to remove a distortion introduced by another component or part of the network. A common example of an amplitude-corrective filter is a gain-flattening filter designed to compensate for the strong wavelength dependence of optical amplifiers. An example of a phase-corrective filter is a dispersion compensator that is used to undo the undesirable effects of dispersion in long-distance communication links, or to remove the chirp from laser pulses.

Almost all optical filters rely on optical interference between two light waves for their filtering behavior. Optical information is carried on waves that oscillate with specific frequencies. Two different waves in the same location can combine to produce a locally higher intensity if their oscillations are in phase (constructive interference), or they can combine to produce a very low or zero intensity if their oscillations are out of phase (destructive interference). The mechanisms used to split the light into parts that can be interfered, and how the interference patterns or outputs are manipulated determine the type of filter. There are a few fundamental configurations or building blocks for optical filters. Each of these is described briefly here and in more detail later.

The Fabry–Perot interferometer (Fig. 1a) utilizes multiple traversals of the same path to produce interference. The signal is split into parts by partially reflecting mirrors at either end of a cavity and reflected back and forth: some parts of the signal traverse the path multiple times before interfering with the other parts of the signal. The fraction of the signal that emerges is determined by interference conditions that depend on the optical path length of the cavity formed by the mirrors, the angle of propagation and also on the reflectivity of the mirrors.

Bulk diffraction gratings act as filters by either reflecting or transmitting (refracting) light in a wavelength-dependent manner. The most common configuration in optical communications is reflection. A reflective diffraction grating consists of closely spaced parallel grooves on a reflective surface (Fig. 1b). The periodic structure of the surface produces an interference between the small fields reflected at each groove that enhances reflections in certain directions, and suppresses them in others. An incident wave is broken up into several orders on reflection from the grating. The directions of the diffracted waves depend on the wavelength of light. Such a device can be used to spatially separate and filter out various wavelengths.

A thin-film stack or dielectric interference filter (Fig. 1c) is a sequence of alternating layers with different refractive

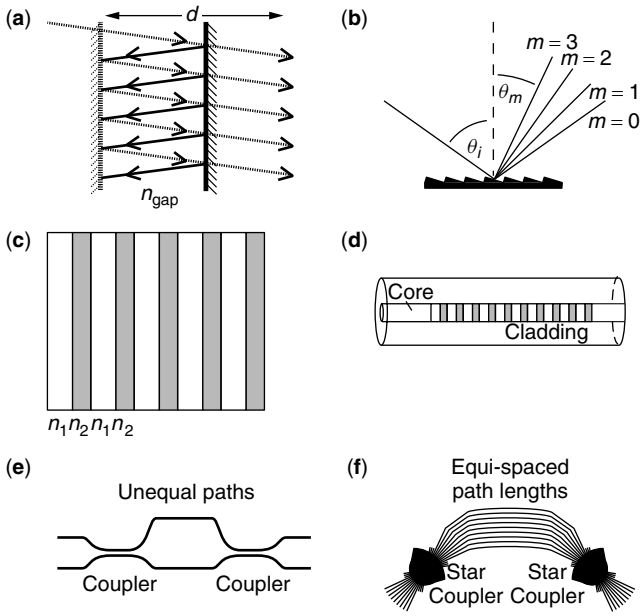


Figure 1. (**a**) A Fabry–Perot interferometer; (**b**) a planar diffraction grating; (**c**) a thin-film interference filter; (**d**) fiber Bragg grating; (**e**) a Mach–Zehnder interferometer; (**f**) an arrayed waveguide grating router.

indices (and sometimes different widths). Small amounts of energy are reflected at each interface between different refractive indices. In a similar way to the Fabry–Perot interferometer, interference between the various reflected waves results in a wavelength-dependent transmission through the stack. The major difference is that the thin-film stack utilizes multiple layers where each interface provides moderate reflectivity; the Fabry–Perot interferometer usually uses a single cavity bounded by a pair of reflecting elements with extremely high reflectivity. (Note that in a Fabry–Perot interferometer the reflecting elements either can be simple metallic mirrors or can themselves be a thin-film stack.)

Waveguide grating filters (Fig. 1d) are essentially dielectric interference filters that exist within the guiding region of an optical waveguide or fiber. Light traveling along the waveguide or fiber interacts with the layered structure and the light is reflected, coupled into different traveling modes of the guide or ejected in a wavelength-dependent manner. If the periodicity of the layers is on the order of the wavelength, the light is usually reflected at the resonant wavelength; these structures are called *Bragg gratings* or *counterpropagating gratings*. Long period gratings, however, usually couple the light to a different mode of the structure traveling in the same direction and are also called *mode-converting gratings*. If the planes of the layers are tilted with respect to the axis of the waveguide or fiber, then light will be coupled out of the guide; these gratings are called side-tap gratings.

The Mach–Zehnder interferometer (Fig. 1e) differs from the Fabry–Perot interferometer and the grating filters in that it acts as a filter by interfering two parts of the signal that have traveled along different paths. The incoming signal is equally split between the two alternate

paths by the input coupler, and on exit the signals in the two paths are recombined by a second coupler. The optical path difference between the two paths determines what fraction of each wavelength appears at each output port. When used as a demultiplexing device, it can be designed to send a particular wavelength completely to one output and another wavelength to the other output.

The arrayed waveguide grating router (Fig. 1f) is a generalization of the Mach–Zehnder interferometer to multiple arms or pathways. Incoming signals can arrive along several input ports. All of these are connected to an input star coupler. The input star coupler splits the signal equally between several pathways that are reconnected at their distant ends to another star coupler. Each pathway differs from its adjacent pathways by a fixed optical path delay. The output star coupler recombines the signals. Interference between the signals determines which output waveguide each wavelength emerges from. The structure is designed so that each wavelength from each input port is shuffled onto a different output port.

1.1. Tunable Filters and Switches

The wavelength(s) of operation of a filter depends (depend) on the optical and geometric properties of the structure. Changing any of these properties will affect the spectral properties of the filter, such as the peak value of transmission, the location of the wavelength at which maximum or minimum transmission occurs, or the extent of the band over which transmission is suppressed. This results in both undesirable effects such as the filter characteristics being temperature- or pressure-sensitive, and desirable effects in that the mechanism can be used to tune the filter.

Various controlling mechanisms (thermal, acoustic, electro-optic, nonlinear) can be used to alter one or more optical characteristics of the filter. If the filter characteristics can be changed sufficiently such that signals are completely diverted from their existing pathways to alternate pathways, the filter can be made to behave as a wavelength-dependent switch. Combining the ability to select wavelengths and to select pathways produces very powerful and useful optical components.

Reconfigurable and adaptive optical components are a current and highly active area of research, as communication networks continually move toward incorporating more flexibility, responsiveness, and intelligence into the optical layer [1,4]. It is far from clear which technologies will emerge as the leaders in this area, though it is likely to involve a hybrid of electronic, optical, micromechanical, and possibly chemical or biological technologies.

2. FILTER EXAMPLES AND APPLICATIONS

2.1. Fabry–Perot Filter

A basic Fabry–Perot filter or interferometer [5,6] consists of an optical cavity between two reflective mirrors. The incident light undergoes multiple reflections from the mirrors at either end of the cavity. The transmitted waves will constructively interfere to produce a maximum when the round-trip optical path length is an integral number of wavelengths:

$$2n_{\text{gap}}\, d \cos\theta = m\lambda \tag{1}$$

This resonance condition depends on the size of the cavity d, the refractive index in the cavity n_{gap}, and the angle of incidence θ.

The transmitted intensity for a cavity with equally reflective mirrors at both ends is given by

$$I_{\text{out}} = I_{\text{in}} \frac{1}{1 + \dfrac{4R}{(1-R)^2} \sin^2\left(2\pi n_{\text{gap}} \dfrac{d}{\lambda} \cos\theta\right)} \tag{2}$$

where R is the reflectivity of each mirror.

The transmission spectrum (Fig. 2) is periodic function of frequency. Each peak is associated with a different order m and the separation of successive peaks or orders is called the *free spectral range* (FSR) and is given by

$$\text{FSR} = \frac{c}{2n_{\text{gap}} d \cos\theta} \tag{3}$$

in frequency units. The sharpness of the transmission peaks is related to the reflectivity of the mirrors. The FWHM Δf (in frequency units) or $\Delta\lambda$ (in wavelength units) is given by the relationship

$$\frac{\Delta f}{f} = \frac{\Delta\lambda}{\lambda} = \frac{1-R}{m\pi\sqrt{R}} \tag{4}$$

The ability of a Fabry–Perot filter to resolve different signals is determined by the ratio of the FSR to the FWHM. This ratio is called the "finesse" $\mathcal{F}$ of a Fabry–Perot filter:

$$\mathcal{F} = \frac{\pi\sqrt{R}}{1-R} \tag{5}$$

Note that the expression for finesse is independent of order m. High finesse or high resolution is achieved for highly reflective mirrors. This expression is for the ideal or maximum value. In any real device the finesse will be smaller because it is limited by imperfections, such as a lossy medium and tilted or nonflat mirrors.

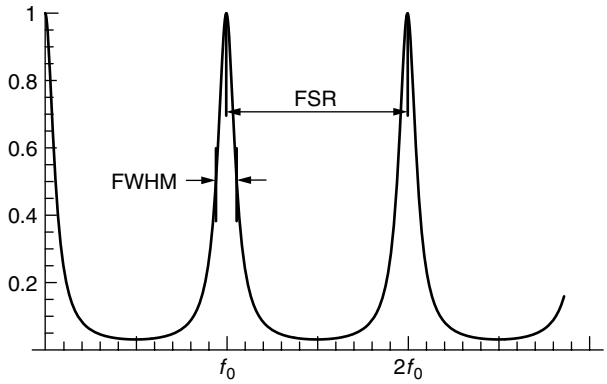


Figure 2. Transmission spectrum for a Fabry–Perot interference filter, depicting the free spectral range (FSR) and the full-width half-maximum (FWHM).

The contrast of the filter is a function of the reflectivity R:

$$\frac{T_{\max}}{T_{\min}} = \frac{(1+R)^2}{(1-R)^2} \tag{6}$$

In all the expressions above, the mirror reflectivity R has been taken to be a constant. In a real device, R is also potentially a function of wavelength and a function of angle, which complicates the analysis of the filtering characteristics. Nevertheless, the operation of the filter remains conceptually the same.

Ideally, one would like very high reflectivities suggesting the use of metallic mirrors. However, this is difficult to achieve in practice without simultaneously introducing excessive loss. Alternatively, these metallic mirrors can be replaced with dielectric stacks that have a broad reflectance peak around the spectral range of interest. The wavelength-dependent R introduced by using dielectric stacks depends on the index difference Δn of the stack. The analysis is not trivial, but as a general rule of thumb it will reduce the FWHM of the filter from the ideal value given above by a factor $\Delta n/\overline{n}$, where $\overline{n}$ is the average index in the stack.

2.1.1. Multipass and Multicavity Cascaded Fabry–Perot Filters. The finesse of a simple Fabry–Perot filter is limited by the reflectivity of the mirrors. Various improvements to the basic filter can be made by modifying the cavity structure [5,6].

A high finesse structure can be made by cascading multiple low finesse structures. There are two approaches. In the *multipass* method, the light passes twice (or multiple) times through the *same* cavity yielding a filter function which is the square (or higher power) of the original filter function. In the *multicavity* method several independent cavities are concatenated and the resulting filter function is the product of the filter functions of the individual cavities. In these configurations it is vital to keep the cavities isolated so that there is no resonating backreflection between the cavities. This can usually be done by slightly misaligning the orientations of each cavity so that spurious reflections will gradually deviate or walk away from the axis of the system.

If the cavities have identical free spectral ranges (FSRs) the cascaded system will have a transmission function with the same FSR but a much narrower FWHM, thus improving the finesse. The free spectral range of the cascaded system can be vastly increased by choosing the cavities to have different FSR (usually in the ratios of different integers). The Vernier principle can be exploited so that the FSR of the cascaded system will be determined by the coincidence of two different orders of the individual cavities (Fig. 3).

2.2. Bulk Diffraction Gratings

Bulk diffraction gratings are surfaces or plates with periodic grooves that can act as either reflective or transmissive structures [1,3]. Light incident on the grooves is split into components that are reflected in various directions depending on the angle of incidence, the wavelength, and the period of the grating. The zero-th order reflected ray is in the direction of specular reflection (equal angles of incidence and reflection); the higher order rays are referred to as diffracted rays and their direction is given below.

The basic diffraction grating equation in reflection is

$$\sin\theta_i + \sin\theta_m = m\frac{\lambda}{\Lambda} \tag{7}$$

where θ_i is the angle of incidence, θ_m is the angle of reflection of the mth-order diffracted ray, Λ is the period of the grating, and λ is the wavelength of the incident light. The fraction of light that ends up in each diffracted order is determined by the detailed shape of the diffraction grooves.

Diffraction gratings can be used in various configurations as wavelength selective elements; the most common is to combine a focusing element (either a lens or a concave mirror) with the grating. One configuration is shown in Fig. 4.

The light emerging from each waveguide hits the focusing mirror, which not only redirects the light onto the grating but also counteracts the diffractive spreading of the light as it exits the waveguide. Each wavelength

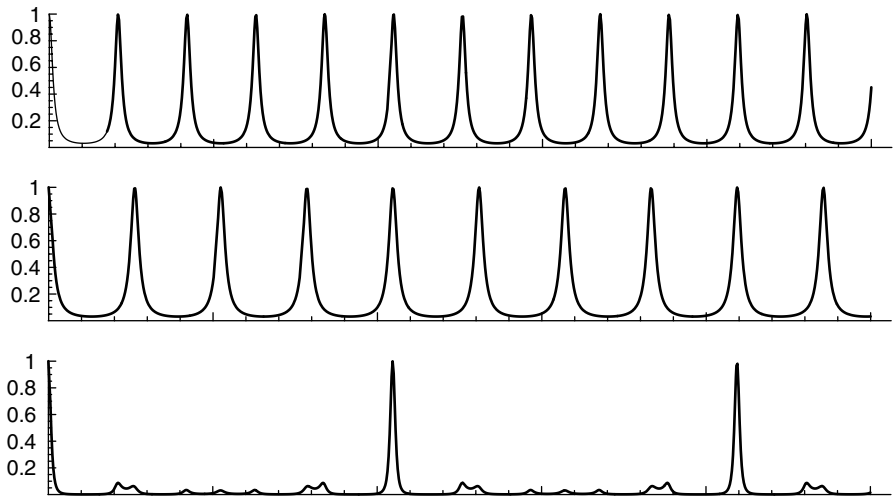


Figure 3. Transmission spectrum for a multicavity cascaded Fabry–Perot interference filter, exploiting the Vernier principle to improve the finesse.

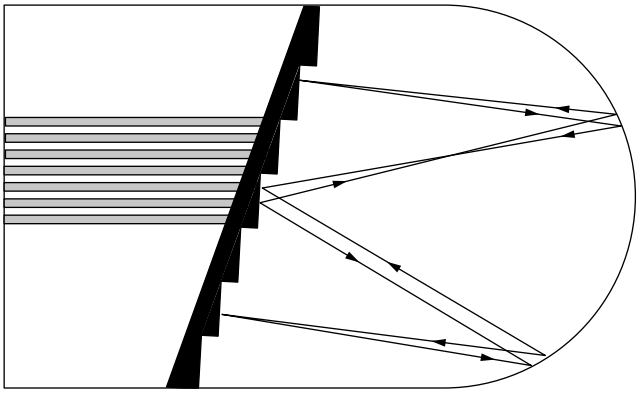


Figure 4. An example of a diffraction grating mounted onto a focusing lens or mirror.

is diffracted by a slightly different angle and therefore is reflected by the mirror to a different position, the subsequent reflections then refocus it onto a different exit waveguide.

2.3. Dielectric Thin-Film Stacks

The first historical observation of thin film interference was in 1817 by Fraunhofer, when he observed that glass with a thin layer of tarnish was less reflective than fresh glass. The simplest antireflection coatings consist of a thin layer of thickness d with an index n_1, which is intermediate between the indices n_0 and n_2 of the materials on either side. In the absence of the antireflective layer, the reflection for normal incidence is $R = r_{2,0}^2$, where

$$r_{i,j} = \frac{n_i - n_j}{n_i + n_j} \tag{8}$$

For an air–glass interface this reflectance is about 4%. In the presence of the thin layer, the reflectance becomes

$$R(\lambda) = \frac{r_{2,1}^2 + r_{1,0}^2 + 2r_{2,1}r_{1,0}\cos\left(\dfrac{4\pi n_1 d}{\lambda}\right)}{1 + r_{2,1}^2 r_{1,0}^2 + 2r_{2,1}r_{1,0}\cos\left(\dfrac{4\pi n_1 d}{\lambda}\right)} \tag{9}$$

When the individual interface reflections are low, this expression can be approximated by its numerator (which is equivalent to ignoring multiple or higher-order reflections). In either case, the minimum reflection occurs when the argument of the cosine function is an odd multiple of π. Thus, the minimum reflection occurs for wavelengths that satisfy $2n_1 d = (m + \frac{1}{2})\lambda$, where m is an integer. When $m = 0$, the thinnest possible antireflective coating is obtained with an optical thickness of $n_1 d = \lambda/4$. This is why these layers are also referred to as *quarter-waveplates*.

The value of the minimum reflection is

$$R_{\min} = \left(\frac{r_{2,1} - r_{1,0}}{1 + r_{2,1}r_{1,0}}\right)^2 \tag{10}$$

This minimum drops to zero if $n_1 = \sqrt{n_0 n_2}$ the geometric mean of the indices on either side.

A stack of thin-film layers of alternating refractive index each one quarter-wavelength thick will produce a very-high-contrast filter, with additional layers producing even greater contrast. A general thin-film filter or stack will consist of many layers of alternating refractive indices. A vast variety of filter characteristics can be designed by varying the layer thicknesses and refractive indices and incorporating more complicated patterns of alternating layers. The filters can be designed to be lowpass filters, highpass filters, and bandpass filters [7]. One general observation is that to obtain sharp cutoff filtering characteristics, the refractive index between the layers needs to be as high as possible. Some of the common materials used for visible light filters and their typical refractive indices are magnesium fluoride (1.39), zinc sulfide (2.35), cryolite (1.35), titanium dioxide (2.3), silicon dioxide (1.46), and various rare-earth oxides. In the infrared, silicon (3.5), germanium (4.0), and tellurium (5.1) can also be used in combinations with low index materials. Countless other oxides and fluorides are also used.

The wide variety of materials and deposition processes available for thin-film filters and the accuracy with which the layer thicknesses can be manufactured make thin-film filters among the most versatile and accurate of filtering devices. Thin-film filters can also be extremely narrow band and accurately manufactured to a precise wavelength; they are the ideal device to use in monitoring the wavelength drift of other devices. For example, a wavelength locker is used to monitor and control the wavelength of a laser source. The locker consists of two cascaded filters with equal bandwidths accurately located on either side of the desired operating wavelength. The optical signals transmitted through this pair of filters are compared and used to provide an electrical feedback signal to compensate for wavelength drift.

A set of thin-film filters can be cascaded to form a WDM demultiplexer (Fig. 5), with each filter either reflecting or transmitting a different specific wavelength channel. The simplest configuration utilizes each filter as a bandpass filter: transmitting a specific wavelength and rejecting all others. The filters can be conveniently deposited on both sides of a transparent dielectric slab, which also provides for accurate parallel alignment of the filters. The output wavelengths can be collimated and launched into fibers with a set of graded-index (GRIN) lenses. Note that since

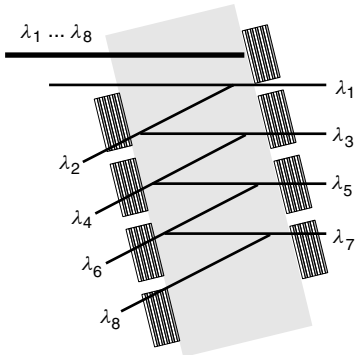


Figure 5. A cascaded set of thin-film filters used to demultiplex eight wavelengths.

all the filters are tilted with respect to the direction of the incident light, the filter layer thicknesses must be designed according to this angle. This type of cascaded filter is suitable for up to about 32 WDM channels.

2.4. Waveguide and Fiber Gratings

Gratings in waveguides and fibers are fundamentally similar to thin-film stacks. Both are structures with periodically varying optical properties [8,9]. The differences arise mainly because thin-film stacks are bulk-optic structures, whereas waveguide gratings are embedded in an existing guiding structure. The embedded gratings thus are restricted to a much smaller range of index differences since they are created by modifying the properties of a single material rather than interleaving different materials. Thus rather than use a moderate number of high contrast layers, embedded gratings use an extremely large number of low contrast layers.

The current importance of Bragg gratings is directly attributable to the photosensitive process whereby permanent index changes can be induced in glass materials using various specific wavelengths of light (mostly in the ultraviolet but also using two photon processes in the visible). An interference pattern is produced by combining two beams of light obtained from a diffraction phase mask or other source (Fig. 6). The periodic pattern of light in turn produces a periodic variation in the induced refractive index change. This is the preferred approach for writing gratings with micron and submicrometer periods. If the period of the grating is longer, then each layer of the grating can in principle be written individually. Those grating structures that can be produced by a holographic process may consist of many (thousands of) periods as opposed to manufactured thin-film stacks that have a much smaller number of layers.

Gratings can be used to selectively reflect light by coupling light from a forward traveling mode of the waveguide or fiber, to a backward-traveling mode. Such gratings are called *Bragg gratings* and have periods comparable to the wavelength of light. Gratings can also be used to selectively couple light between *different* modes traveling in the same direction. Such gratings are called *mode-converting* or *long-period gratings* and can have periods from several micrometers to centimeters.

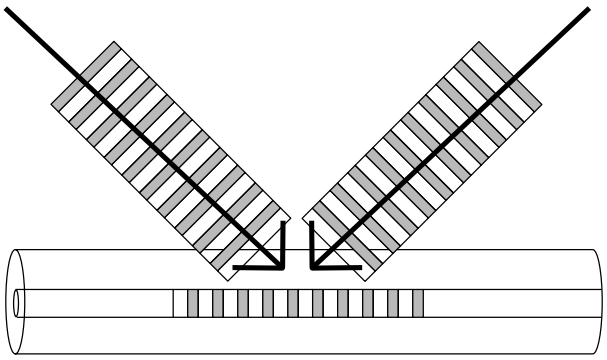


Figure 6. Two interfering beams of light used to produce a grating.

The relationship between the resonant wavelength (or Bragg wavelength) and the period of the gratings involves the effective refractive indices of the modes of the underlying waveguide. In any waveguide, each mode has its own characteristic phase velocity, which determines an effective refractive index (intermediate between the smallest and largest refractive indices occurring inside the waveguide structure). For a grating that couples between modes A and B the resonance condition is

$$\lambda_0 = (n_A \pm n_B)\Lambda \tag{11}$$

where Λ is the period of the grating, n_A and n_B are the effective indices of the modes and λ_0 is the free-space wavelength at which the coupling is most efficient. The positive sign is used for Bragg gratings (where the modes are traveling in opposite directions), and the negative sign is used for gratings that couple modes traveling in the same direction.

The actual efficiency of the grating and how it varies as the wavelength departs from the resonant value depend on other properties of the grating such as its length and the depth of modulation of the refractive index. This is explored briefly later in the section on coupled mode theory. In general, for simple gratings, Bragg gratings operate over a narrow range of wavelengths (a few nanometers) and long-period gratings operate over many tens of nanometers.

A wide variety of spectral characteristics can be obtained from gratings for a diverse range of applications [8,9]. Common applications of Bragg gratings are in optical add drop multiplexers (OADMs). Gratings can be easily designed to strongly reflect over a narrow range of wavelengths, referred to as the bandgap or reflection band of the grating. Unfortunately, most WDM applications of filters require a bandpass rather than a bandreject functionality and so various configurations have been developed to exploit reflective gratings.

Two different configurations are shown in Fig. 7. In the first configuration, the grating is located between two 3-port circulators. All incoming wavelength channels are sent by the first circulator toward the grating. All wavelength channels except one are passed by the grating and then sent by the second circulator back out to the network.

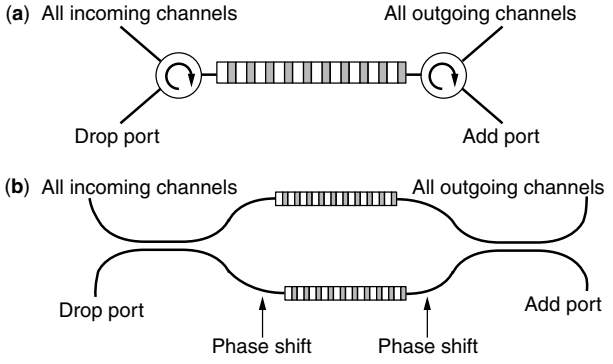


Figure 7. (**a**) Optical add-drop multiplexer formed by a grating and two circulators; (**b**) alternative configuration with two couplers.

The grating is designed to reflect the wavelengths corresponding to one of the channels: the drop channel. These wavelengths are reflected by the grating and after passing back through the first circulator are delivered to the drop port. Likewise, signals arriving at the add port are passed by the second circulator toward the grating. The add channel is reflected by the grating, and then, after passing back through the second circulator, these wavelengths join the other channels passing out to the network.

The high cost of circulators has stimulated alternative designs. The second configuration requires a pair of *identical* gratings located between two couplers. The drop channel wavelength is reflected by both gratings simultaneously, but picks up an extra π phase shift in one arm. As it passes back through the coupler, it is recombined onto the other port of the coupler, because of the phase shift. The analogous thing happens on the other side with the drop channel.

Another important application of gratings is in gain flattening. The output of an erbium-doped fiber amplifier (EDFA) varies strongly as a function of wavelength over its band of operation (Fig. 8). When a signal passes through many such amplifiers, the nonuniformity is accentuated, and this severely limits the usable bandwidth of the system. Gratings having a filter profile that is the opposite of the gain spectrum of the EDFA can be used to flatten the profile and thus extend the usable bandwidth. Various types of long-period gratings have been successfully used to modify the profile over a range of tens of nanometers.

The final important application of gratings considered here is the dispersion compensator. Signals traveling over long distances in standard telecommunications fiber near 1.55 μm experience dispersion-induced broadening since the longer wavelengths travel slightly more slowly than shorter wavelengths. This group velocity dispersion will lead to signal degradation. A chirped Bragg grating is one where the period of the grating varies along its length. If the grating period is longer at the front than at the back, the shorter wavelengths will travel further into the grating before being reflected. This introduces a wavelength-dependent delay on reflection that can be used to compensate for the delay induced by dispersion (see Fig. 9). The dispersive properties of gratings are a very important area of investigation [10].

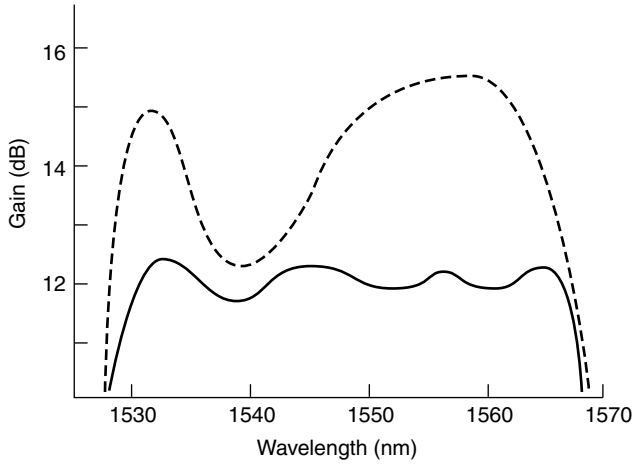


Figure 8. Output spectrum of a typical EDFA before (dashed) and after (solid) passing through a gain-flattening filter.

Figure 9. Dispersion-compensating Bragg grating.

A major problem associated with gratings is temperature dependence. Not only does the grating physically expand with temperature causing the period to shift, but a more important effect is the temperature-dependent refractive index of most glasses, which also results in a shift in the resonant or Bragg wavelength. Temperature compensation in grating packaging is a critical component of grating technology.

2.4.1. Coupled-Mode Theory. Coupled-mode theory is used extensively for the modeling and analysis of gratings. The coupling strength of the grating is encapsulated in a parameter κ, which is proportional to the index modulation. The frequency of the incoming signal is described by a detuning that is proportional to the frequency difference between the signal and the resonant or Bragg frequency.

More precisely, the coupling strength is defined by

$$\kappa = \eta \frac{\pi}{\lambda} \Delta n \tag{12}$$

where Δn is the index modulation and η is an efficiency factor describing how well the grating overlaps with the transverse intensity profile (or modal profile) of the light traveling along the waveguide; the detuning is defined by

$$\delta = (n_A \pm n_B)\frac{\pi}{\lambda} - \frac{\pi}{\Lambda} \tag{13}$$

where, as before, the positive sign is used when the modes are counterpropagating and the negative sign when they are copropagating.

A general rule of thumb for all gratings is that coupling is very efficient if the detuning δ is smaller than the grating strength κ and is very inefficient when the detuning is larger than κ. For Bragg gratings the detuning range between $\pm\kappa$ is also referred to as a *bandgap*.

The coupled-mode equations are the fundamental system of equations used to describe both Bragg (copropagating) gratings and long-period (counterpropagating) gratings [8,9]. The first mode (usually forward-traveling) has an amplitude that varies as $u(z)$; the second mode (either forward- or backward-traveling) has an amplitude varying as $v(z)$. The coupling of energy by the grating is represented by a pair of simple differential equations connecting these two amplitudes. For uniform gratings the equations are

$$iu'(z) + \delta u(z) + \kappa v(z) = 0 \tag{14}$$

$$-iv'(z) + \delta v(z) \pm \kappa u(z) = 0 \tag{15}$$

where the positive sign is used with counterpropagating modes and the negative sign is used with copropagating modes.

A typical spectrum for a uniform Bragg grating is shown in Fig. 10. The peak reflectivity is given by

$$R = \tan h^2(\kappa L) \tag{16}$$

where L is the length of the grating. The longer the grating, the stronger the reflection. The bandwidth of the spectrum is directly proportional to κ and can be represented in

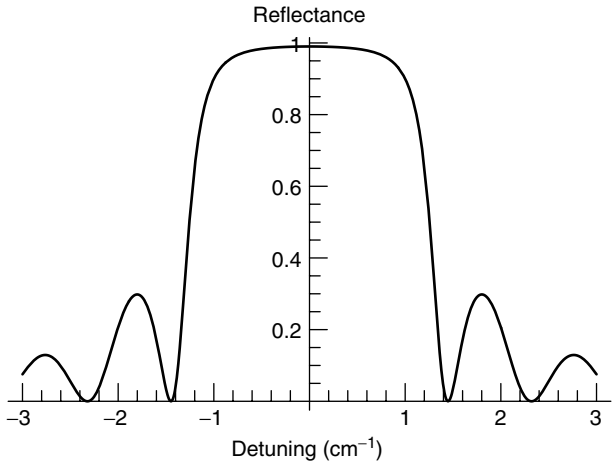


Figure 10. Reflection spectrum for a Bragg (counterpropagating) grating with a total grating strength $\kappa L = 3$.

various ways. As a rough rule

$$\frac{\Delta\lambda}{\lambda} \sim \frac{\Delta n}{n} \tag{17}$$

Very short gratings will have an apparent bandwidth wider than this, but as the grating becomes longer, the spectrum will stabilize to this fixed bandwidth. Making the grating extremely long will *not* narrow the reflection band any further.

For copropagating (or long period) gratings, it is still true that the most efficient energy coupling occurs at zero detuning. However, since both modes continue to travel in the same direction, if the grating is long enough, the coupling process will start to couple energy back into the original mode again. This phenomenon is called *overcoupling*. Figure 11 shows three typical spectra for a copropagating grating demonstrating ideally coupled and overcoupled gratings. The ideal condition for 100% coupling is first achieved for $\kappa L = \pi/2$.

2.5. Acoustooptic Filters

Acoustooptic filters exploit the interaction of light and sound in materials that are photoelastic [1,4]. The periodic compressions and expansions in the material produced by the presence of the sound wave result in corresponding variations in the refractive index via the photoelastic coefficient. A commonly used material with a high photoelastic coefficient is lead molybdate, $PbMoO_4$.

The period Λ of the index modulation produced by the photoelastic effect is equal to the wavelength of the

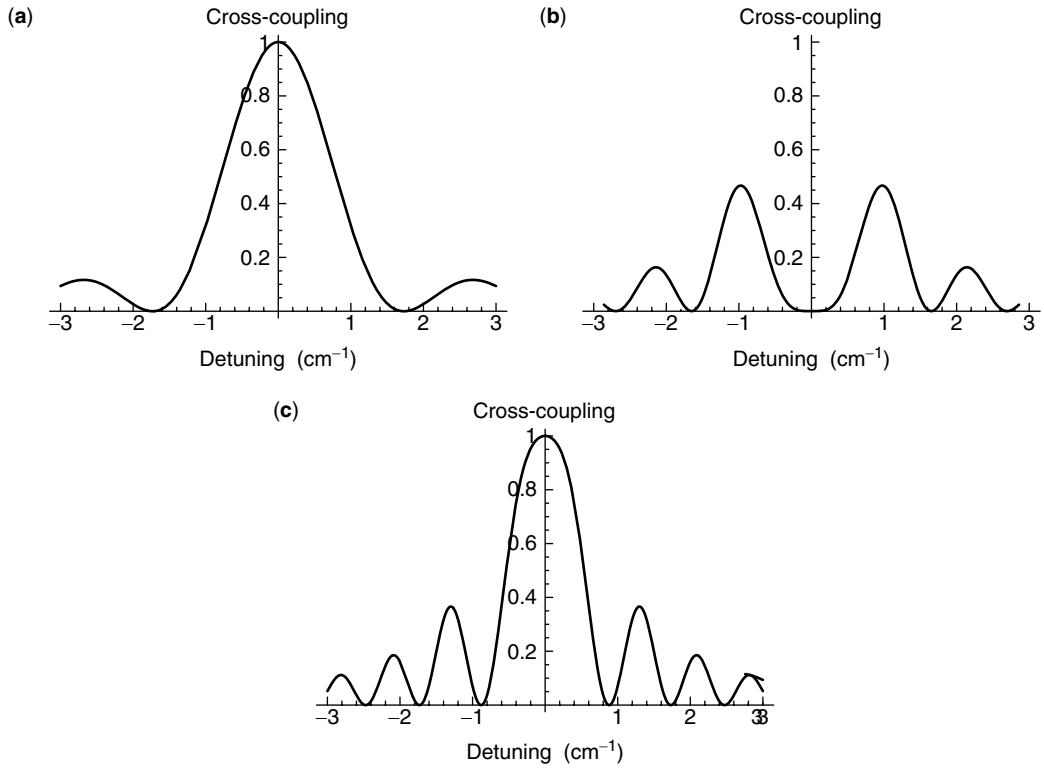


Figure 11. Cross-coupling spectra for copropagating gratings with three different values of total grating strength: (**a**) $\kappa L = \pi/2$; (**b**) $\kappa L = \pi$; (**c**) $\kappa L = 3\pi/2$.

sound wave in the material. Sound velocities in typical materials are on the order of several thousand meters per second. Thus acoustic frequencies of around 100 MHz will produce gratings with periods of several tens of micrometers. A common configuration is to use such gratings to couple energy between lightwaves traveling in the same direction but experiencing different refractive indices. The different refractive indices are obtained by using a birefringent material and exploiting the different refractive indices (ordinary n_o and extraordinary n_e) for the two polarizations. The wavelength at which maximum coupling of energy from one polarization to the other occurs is given by the condition

$$\lambda = (n_o - n_e)\Lambda \tag{18}$$

The bandwidth over which effective coupling occurs is roughly given by the relationship

$$\frac{\Delta\lambda}{\lambda} \sim \frac{\Lambda}{L_{\text{int}}} \tag{19}$$

where the interaction length L_{int} is the distance over which the acoustic and optical waves overlap and interact.

Acoustic gratings share many properties with waveguide gratings and thin-film stacks and can be analyzed by the same techniques. However, the most important difference is that the acoustic grating is a transient phenomenon, and can be easily controlled or modified by changing the properties of the acoustic wave, thus leading to various tunable filter configurations.

2.6. Mach–Zehnder Interferometer

The basic principle behind the Mach–Zehnder (MZ) interferometer is the interference of two parts of a signal that have traversed different optical paths [11]. In one of the simplest configurations (Fig. 1e) two waveguide arms of different optical lengths are connected by two 3-dB couplers. For simplicity, any wavelength-dependent properties of the couplers themselves are ignored (at least over the wavelength interval of interest). The light is split equally at the first coupler and recombined at the second coupler. The interference is produced by the phase difference between the waves traversing the two arms of the interferometer. The phase difference is given by

$$\Delta\phi = \frac{2\pi}{c} f n_{\text{eff}} \Delta L = \frac{2\pi n_{\text{eff}}}{\lambda} \Delta L \tag{20}$$

The intensities obtained from the two ports, respectively, are given by

$$I_1 = I_{\text{in}} \cos^2 \frac{\Delta\phi}{2} \tag{21a}$$

$$I_2 = I_{\text{in}} \sin^2 \frac{\Delta\phi}{2} \tag{21b}$$

Thus, interference between the two arms leads to a difference in power between the outputs of the second coupler.

The filtering characteristics are periodic in frequency and the channel spacing (also called *free-spectral range*) of this device is

$$\Delta f = \frac{c}{2n_{\text{eff}} \Delta L} \tag{22}$$

In more realistic devices, the perfect periodicity of the transmission spectrum will be modulated by the wavelength-dependent properties of the 3-dB couplers (which in turn depend on how the couplers are made).

If we contrive to have constructive interference for a specified frequency while having destructive interference for a second specified frequency, this will determine the length. For a 1.3/1.55-μm channel splitter, the length required is $L = 2.78$ μm (assuming a silica waveguide with $n_{\text{eff}} = 1.45$). The filtering characteristics are shown in Fig. 12. On the other hand, to create a device that separates 100-GHz channels (interleaving them) would require a length of $L = 1$ mm.

The filter can also be used in reverse as a multiplexer for combining wavelengths. For example, to combine a 980-nm-pump wavelength with a 1550-nm signal, the appropriate length would be $L = 0.92$ μm (for a pump at 1.48 μm the length becomes $L = 11.3$ μm). However, it is more common for this operation to be done by exploiting the wavelength-dependent coupling of a single simple directional coupler. Such a wavelength-dependent coupler can also be used to separate two wavelengths.

A cascaded series of MZ filters (sometimes called an MZ "chain") can be used to systematically separate or demultiplex a full-WDM channel set. Consider a set of equally spaced frequencies. The first MZ filter is designed to separate all the even channels from all the odd channels. Thus each output arm of the first device now carries a set of equally spaced frequencies with a spacing twice that of the original signal set. A second MZ filter that has a pathlength difference ΔL half that of the original will subsequently filter out every other channel in this reduced set and so on. Thus, for example, a cascaded chain of MZ filters 7 deep could demultiplex a 128-channel system.

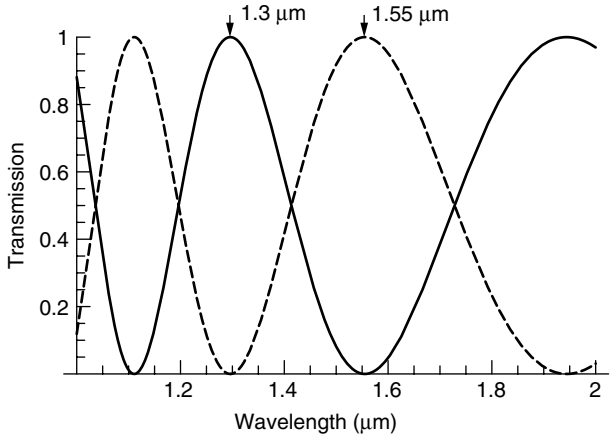


Figure 12. Filtering characteristics of the two output ports as functions of wavelength. If shown versus frequency, the characteristics would appear periodic.

2.7. Arrayed Waveguide Grating Router

The arrayed waveguide grating router (AWGR) can be regarded as a generalization of the Mach–Zehnder filter to multiple ports [3,4]. The input and output couplers are replaced by star couplers that have the property that the light entering at any port is split equally between all output ports. (Information about which specific input port the light came from is retained or encoded into the relative phases of the split signals.) If the multiple paths connecting the two star couplers were all of identical optical lengths (the relative phases would be preserved), then the second star coupler would just undo or reverse the action of the first star coupler; light entering a specific port on the left emerges from the corresponding port on the right (for all wavelengths). When the optical pathlengths of the arms are different, the signals arriving at the second star coupler will have different phases and interference will direct the output to a different port (which port will also depend on wavelength).

A useful analogy to understand the AWGR is to replace the star couplers with lenses and the array of arms with a triangular dispersive prism as in Fig. 13. The different input ports correspond to different point sources A, B, and so on. in the focal plane of the input lens. Light from each of these point sources after passing through the lens is transformed into *plane waves* traveling at different angles, each angle corresponding to a different input source. In the absence of any intervening structures, plane waves hitting the output lens are focused onto its focal plane, with the position of the image determined by the incident angle. Thus the arrangement of the output images corresponds precisely to the input sources (apart from a simple overall inversion).

The triangular prism refracts each incoming wave changing its direction, thus altering the location at which the output image is formed. Furthermore, because the amount of refraction will vary for different wavelengths, each wavelength will form an output image in a different location.

Using the notation $\lambda_i^{(j)}$ to indicate a channel with wavelength λ_i entering through port j, where $i = 0, 1, \ldots, N-1$ and $j = 0, 1, \ldots, N-1$, then such a signal emerges from port $j + i$(moduloN), (i.e., if $i + j$ is larger than $N - 1$, the port numbers wrap around.) The 4×4 example is shown explicitly in Fig. 14. The wraparound effect is obtained by ensuring that the frequency spacing between adjacent channels Δf and the optical pathlength difference $n\Delta L$ between adjacent arms satisfy $N \times \Delta f \times n\Delta L = 2\pi c$.

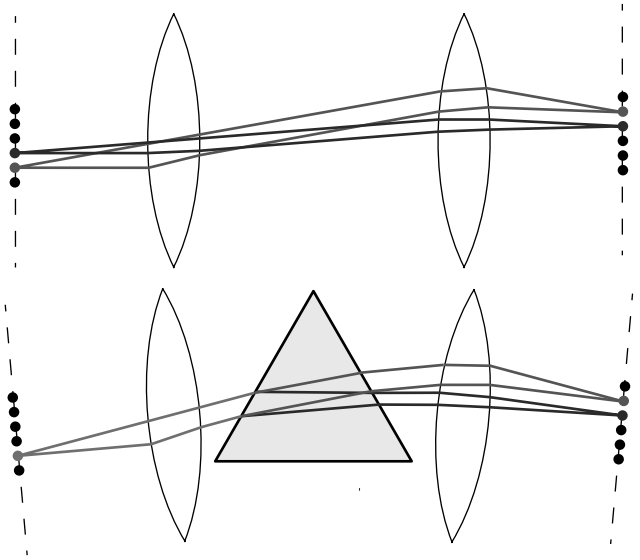


Figure 13. Analogy of an AWGR using two lenses and a triangular prism.

3. FILTER CHARACTERISTICS

Several important parameters characterize the spectrum of a filter, especially WDM filters designed for channel selection. The parameters are shown for a typical spectral profile in Fig. 15.

The peak wavelength is as the name suggests: the wavelength with the least loss. For an asymmetric spectrum, this will be different from the *center wavelength*, which is defined as the average of the upper and lower cut wavelengths or limits of the passband. Flexibility exists in the definition of these limiting wavelengths, as they are the wavelengths on either side of the peak for which the spectrum first falls to some predetermined level in decibels. For example, the 0.5- and 3-dB bandwidths are shown in Fig. 15. If the spectrum is significantly asymmetric, the bandwidth and centre wavelength will depend on the choice of level.

The isolation is given by the largest transmission levels (lowest loss) in the adjacent WDM channels. It is not necessary that this extreme value occur at the edge of the adjacent channels, nor that it even occur for the immediately adjacent channel. The figure of merit for

Figure 14. Input port and wavelength redistribution for a 4×4 AWGR.

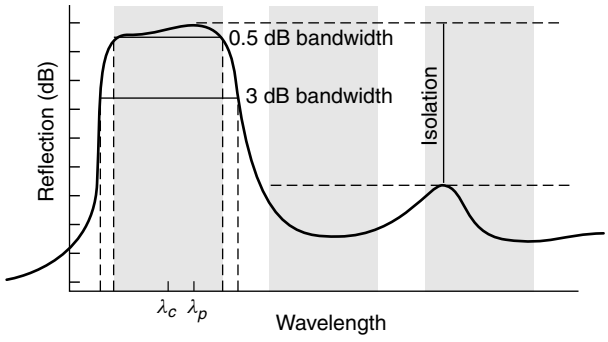


Figure 15. Typical WDM filter profile showing both the peak and central wavelengths, the bandwidth at various decibel levels, and the channel isolation or crosstalk.

a WDM filter can be defined by the ratio of utilizable bandwidth to the channel separation.

4. EVALUATION

Each filter technology or configuration has its own specific merits and disadvantages; the various designs are all competing for use in commercial optical network systems. Some of the important criteria are precision, temperature stability, low loss, low polarization dependence, manufacturing cost, packaging and pigtailing (i.e., connecting fibers to chip-based devices), scalability to large numbers of wavelength channels, isolation and crosstalk between channels, and design flexibility.

In terms of wavelength precision and stability, the thin-film filters are extremely attractive components for optical networks. The deposition process yields very accurate layers, and the variety of materials available make it easy to design for low loss and minimal temperature dependence [4] (as low as 5×10^{-1} nm/°C). The polarization dependence for these devices can also be kept low. However, scaling to larger numbers of channels usually requires the cascading of devices, which could impact both cost and overall loss.

An advantage of diffraction gratings is that the insertion loss is independent of the number of channels, making these devices more attractive for high-channel-count systems. Unfortunately, one problem with diffraction gratings is their polarization-dependence. These devices also do not fare too well on the cost criterion; the diffraction grating/concave reflector configuration is a bulk device that is not easy to fabricate.

Devices that have the best polarization-independent performance are the fiber Bragg grating filters. Passive temperature stabilization is possible for fiber gratings, giving stability similar to that of thin-film filters [4]. The holographic fabrication technique also provides this class of devices with an unrivaled ability to make unusual and customized filter responses by simply varying the chirp and apodization profiles of the grating. This technology does not excel in scaling with the number of wavelengths. It is possible to overlay a limited number of gratings with different center wavelengths, or to concatenate several gratings within the same length of fiber; however, most multiplexing schemes still require some form of cascading, which introduces the additional overheads of couplers, isolators, and circulators.

Certainly, arrayed waveguide grating routers appear to have the advantage of scalability, at least to moderate numbers of channels (from 16 to 64). However, because this is a planar or chip-based technology, polarization dependent loss may be a problem. A number of strategies for reducing the polarization dependence are available [3]; the simplest is to introduce a polarization-retarding plate half-way across the array waveguide region. The biggest disadvantage of the AWGR is the packaging and pigtailing cost of connecting fibers to the device.

The demand for greater bandwidth will continue to drive the development of components for WDM networks. There is a growing shift toward more intelligence within the optical layer of such networks, suggesting a trend toward devices with multiple functions. For example, early WDM filters concentrated mostly on filter shape, whereas later designs will also look closely at the implications of dispersive effects. Filters will be both selective and corrective at the same time, as well as combining other attributes such as switching and tunability.

The need to reduce packaging costs will also favor devices that can easily be integrated, and the incorporation of light sources, modulators, filters, and detectors onto the same substrate may be one possibility to achieve a scaling in production.

Although it has been possible to identify which attributes will become more important in future optical filter designs, it is still not possible to tell which of the existing technologies will best evolve to fit the needs of the next generation of optical networks. Almost certainly a diversity of designs and technologies will continue to exist, as no single solution will be able to meet all the needs of optical networking.

BIOGRAPHY

Leon Poladian received his B.Sc. degree in 1986 and a Ph.D. degree in theoretical physics in 1990 from the University of Sydney, Australia. His thesis was on the optical properties of periodic structures. He joined the Optical Sciences Centre at the Australian National University in 1990 as a postdoctoral fellow working on nonlinear fibre couplers and spatial solitons. Since 1992, he has been at the Optical Fibre Technology Centre at the University of Sydney, first as a Queen Elizabeth II fellow, then an Australian Research Council senior research fellow and currently as an Australian professorial fellow. Dr. Poladian has published over 80 journal and conference papers and holds five patents in the areas of fiber design and grating fabrication. His areas of interest are computational algorithms for novel fiber design; grating design, fabrication and characterization, and the optical properties of periodic and almost-periodic photonic structures in one, two, and three dimensions. Dr. Poladian also holds a graduate diploma in education from the University of New England, Armidale, Australia.

BIBLIOGRAPHY

1. J. Paul and E. Green, *Fiber Optic Networks*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
2. K. Nosu, *Optical FDM Network Technologies*, Artech House, Boston, 1997.
3. H. J. R. Dutton, *Understanding Optical Communications*, Prentice-Hall, Englewood Cliffs, NJ, 1998.
4. R. Ramaswami and K. Sivarajan, *Optical Networks: A Practical Perspective*, Morgan Kaufman, San Francisco, 1998.
5. W. H. Steel, *Interferometry*, Cambridge Univ. Press, Cambridge, UK, 1983.
6. J. M. Vaughan, *The Fabry-Perot Interferometer*, Adam Hilger, Bristol, UK, 1989.
7. H. A. Macleod, *Thin-Film Optical Filters*, Adam Hilger, London, 1969.

8. T. Erdogan, Fiber grating spectra, *J. Lightwave Technol.* **15**: 1277–1294 (1997).

9. R. Kashyap, *Fiber Bragg Gratings*, Academic Press, San Diego, 1999.

10. G. Lenz, B. J. Eggleton, C. R. Giles, C. K. Madsen, and R. E. Slusher, Dispersive properties of optical filters for WDM systems, *IEEE J. Quant. Electron.* **34**: 1390–1402 (Aug. 1998).

11. P. Hariharan, *Optical Interferometry*, Academic Press, Sydney, 1985.

OPTICAL MEMORIES

MICHAEL RUANE
Boston University
Boston, Massachusetts

1. INTRODUCTION

An optical memory is any system that stores and retrieves digital information using optical methods. Optical memories are mass storage devices, competing directly with magnetic hard drives and magnetic tape. Their high capacity, low cost per megabyte, reliability, and removability have made optical memories the preferred mass storage solution for many applications. They are standard components in personal computers and workstations, and support important consumer electronics. Digital versatile disc (DVD) players are now making significant inroads into the VHS tape market, and rewritable DVD (DVD-RW) is a candidate for the local storage of movies and multimedia distributed over the Internet. Optical memory is a natural component of an all-optical system for data retrieval, transmission, and storage, and will play an important role in advanced communications systems.

As of 2001, global production of content is expected to require 1–2 exabytes or roughly 1.5 billion gigabytes of storage. This is approximately 250 MB (megabytes) per person for every person on earth. This content in print, film, magnetic, and optical forms exceeds the production of content for all history before this year [1]! Consumer high bandwidth applications drive the need for inexpensive, removable memory, while high-performance military, industrial, and enterprise processing systems are investigating advanced optoelectronic devices and optical interconnections that will naturally interface with optical memories. Optical jukeboxes, for example, support enterprise-level storage-area networks (SAN), storing multiple terabytes in one system.

Sony and Phillips first developed optical media and players for the distribution of digital audio in the 1970s, building on laserdisc technology. The compact disc–digital audio (CD-DA or simply CD) was standardized in 1980 to provide a high-fidelity alternative to the conventional LP (long-playing) vinyl record. The computer industry quickly recognized that the CD, configured with stronger error correction as a data read-only memory (CD-ROM), allowed inexpensive distribution of large volumes of data [one CD-ROM holds about 450 HD (high-density) floppy disks]. In 1988 write-once optical memories CD-recordable (CD-R) allowed users to create small numbers of their own discs for storage, testing, or distribution. Fully rewritable systems [CD-rewritable (CD-RW)] followed in 1996, enabling removable optical RAM. Early CD-RW systems were much slower than comparable magnetic devices, but these problems have been largely overcome through higher rotation speeds and the availability of data compression.

DVD technology arrived in the late 1990s. DVD increased optical memory capacity and data transfer speed, making read-only optical memories suitable for full motion video and large data sets, such as high-resolution images. The 10-millionth DVD videoplayer was sold $3\frac{1}{2}$ years after introduction; it took 7 years to ship the 10-millionth CD audio player. Newer PCs increasingly have DVD drives, and DVD-R and DVD-RW units now entering the marketplace are providing high-capacity write-once and rewritable optical memories. Still in the laboratory are optical systems that further extend serial, disc-based optical memories, and new page-oriented optical systems based on holography. Such advanced systems seek capacities above 125 GB/disc and data transfer rates of 25 MBps (megabytes per second), about 25 times that of DVD. Figure 1 summarizes the growth of optical memory capacity.

Despite this success, optical memories face vigorous, application-dependent competition. Storage-area networks both compete with local removable optical memories and create demand for increased server memory. Magnetic tape and hard-disc systems promise significant improvements in capacity, data transfer rate, and access times [2]. Innovative data compression will also change the competitive memory scene, allowing tradeoffs among processor speeds, data transfer rate, and memory.

Optical memories are a specialized communications link (Fig. 2). The "transmitter" of an optical memory encodes, modulates, and writes digital data to the media "channel." Every media channel will introduce characteristic distortion, errors, and noise whose characteristics are determined by the interaction of the optical writing and reading systems with the storage medium. The "receiver" actively probes the channel to create a read data signal, which is detected and decoded to recreate the user's stored data. An important distinction is that stored data can remain "in the channel" indefinitely through the physical modification of the media during writing. The read system recovers stored data when a read beam interacts with the modified media, and experiences modulation of phase, amplitude, polarization, or frequency. The channel media can age, suffer abuse, or simply deteriorate from many cycles of the storage process. Channel characteristics also depend on the performance of control systems for maintaining rotation speed, tracking and focusing.

2. DISC-BASED SERIAL MEMORIES

This section discusses the general characteristics of available optical memories [3]. Commercial optical memories store data as an encoded serial binary bitstream on a

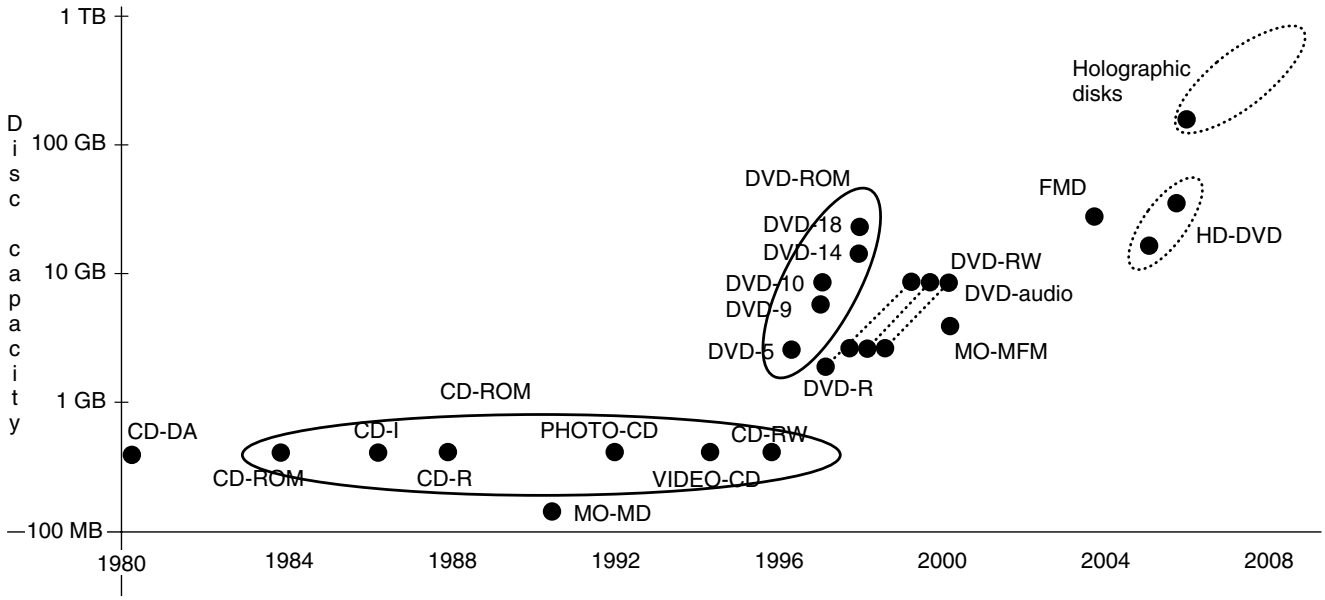


Figure 1. Evolution of optical memory capacity. Media and head constraints dictated initial CD-ROM capacity while different application formats evolved. DVD laser improvements enabled most of the 1990s increase in single layer systems, while disc capacities grew with multilayers. Continued growth will come from blue lasers in high-density DVD, fluorescent multilayers (FMD), and holographic discs. Other approaches are less established.

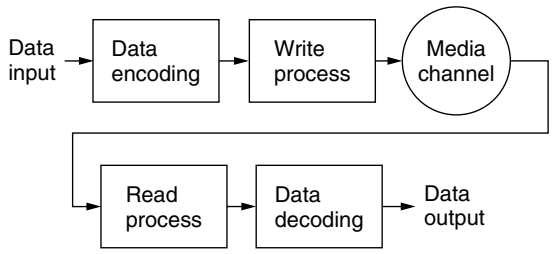


Figure 2. Block diagram of an optical memory. Optical memories are a form of communications link where the data reside in the media channel. User data undergo error coding and is organized into blocks and directories to produce a datastream that drives the write process. The "media channel" stores the data as an imperfect modulation of some property of the media. Readout recovers the media data, but with noise, bursty data loss, crosstalk, and other problems. Data decoding removes overhead, performs error checking, and regenerates the original data.

disc with one or more storage layers, accessible from one or both sides. Disc-based memories are standardized to a 120-mm-diameter disc, although minidiscs (80 mm), card discs, and other formats are specified under published standards.

CD-ROM and DVD-ROM media are mastered and injection molded. First a laser records data as exposed marks in photoresist spun on a polished glass substrate. When developed, the photoresist has holes where the data were written. Next, a conductive conformal nickel layer is deposited and thickened by electroforming to become the "father" stamper, from which "mother" and "daughter" stampers are produced. A daughter stamper is mounted in an injection molding system which is filled with molten polycarbonate under high pressure to form the final CD or DVD substrate. Injection molding takes about 10 s per disc; about 3000 CDs can be produced from one stamper. A reflective metallization layer, usually aluminum, is sputtered onto the molded substrate. Finally, a protective lacquer coating and screen printing or labels are applied.

Recordable and rewritable discs are more complex, with a mastered spiral tracking groove that usually has header marks, focusing marks, and even laser power calibration areas. Grooves sometimes have a deliberate mechanical timing wobble in their walls. An active layer stores the data by absorbing energy from the write beam and changing in some manner [4]. A reflective layer and possibly a complex optical stack enhance the active layer.

CD-ROM standards proliferated as optical memories attracted new applications that required better error control, storage of mixed media, management of multiple sessions, and extended capacities. All CD-ROMs have physical leadin and leadout areas to identify the start and end of data, and use physical data sectors of 2352 bytes, read at 75 sectors per second in a single-speed ($1\times$) drive. Different CD-ROM standards distinguish how the 2352 bytes are allocated to data, synchronization, headers, error detection, and error correction. Some standards require that entire discs must be written at one time, while others allow multiple sessions. Generally, data must be organized to meet all disc and data sector formatting requirements, and recorded without gaps (buffer underflow).

Data must be error encoded and framed to be reliable. Disc mastering has unavoidable defect rates of 10^{-4} or 10^{-5} while surface contamination often destroys or

obscures many adjacent marks. To combat these burst errors, interleaving and Reed–Solomon encoding are used. In CD audio, for example, stereo 16-bit samples are taken at 44.1 kHz, making four 8-bit symbols. A shortened Reed–Solomon (28,24) code operates on a frame of six stereo samples, or 24 8-bit symbols. The 28 output symbols are interleaved and encoded by a (32,28) Reed–Solomon code. Those 32 symbols are regrouped in even–odd groupings, extended by 8 bits for control and display, and then modulation encoded as "eight-to-fourteen modulation" (EFM), with 3 merge bits. This run-length encoding makes optical pickup and timing more reliable. An additional 27 synchronization and merging bits are then added, such that the initial frame of 192 user bits is expanded to 588 encoded bits; 32 such frames form one physical sector.

Encoded and run-length modulated data drive the laser that exposes the photoresist or modifies the active layer in a recordable medium. When read, the medium modulates the amplitude, polarization, or phase of the read beam. Frequency modulation, while possible, is not yet competitive. CD-ROM and CD-R modulate net reflectivity at the disc surface, changing read beam intensity by controlling diffraction from the physical relief of the mastered data pits or controlling the reflectivity of recordable dye materials. CD-RW switches a phase change material between crystalline and amorphous states, depending on its temperature rise under laser heating and subsequent cooling. Having different reflectivities, these amorphous and crystalline states are read as amplitude modulated data. In magnetooptical discs, thermomagnetic laser writing modifies the magnetic state of certain amorphous thin films. Read beam polarization is rotated by the magnetic state, recreating the stored data modulation.

Optical memory access time, the sum of track seek time, drive settling time, and track latency, is relatively slow. An optical head is complex (Fig. 3) and slow to accelerate. Constant-linear-velocity (CLV) drives must adjust rotation speed for each seek. For a 1× DVD, with constant linear velocity of 3.8 m/s, rpm (revolutions per minute) changes from 574 rpm at the outermost track to 1528 rpm at the innermost track. Access times are not greatly improved by higher disc speeds, but greater linear velocity improves data transfer rate. CD-ROMs are read at about 1.41 Mbps (megabits per second) (1×) while 1× DVDs have a user rate of 11.08 Mbps. Optical memories using zoned linear velocity (ZLV) eliminate settling time during localized head movements. Some advanced DVD drives, mimicking magnetic drives, use constant angular velocity (CAV).

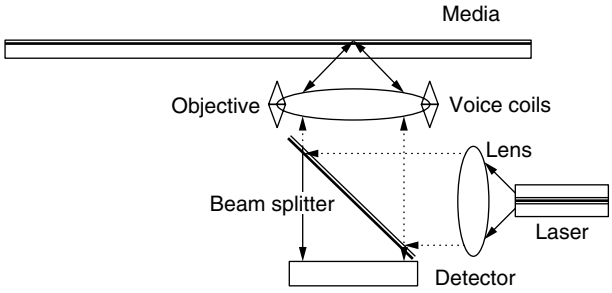


Figure 3. Schematic of a CD head. The laser diode emits elliptical light that is circularized and collimated by a lens. A thin plate directs the beam to the objective lens for focusing through the substrate and onto the pits, just below the lacquer layer. Reflected light, modulated by the effective surface reflectivity, is directed to a detector which reads the data and also senses tracking and focusing servo signals. Voice coils on the objective allow fine pitch tracking and focusing while the entire head moves for coarse tracking.

Optical memories have good lifetime and reliability. Heads fly millimeters above the media, avoiding catastrophic head crashes, while accelerated life testing predicts a shelf life of decades for aluminum-sputtered (silver) ROM media, and longer for gold media. Recordable and rewritable media have projected shelf lives of about 100 years. Mechanical handling damage is a concern, but most contamination is a minor problem because surface contaminants are out of focus. CDs can correct up to about 500 bytes (a 2.4-mm scratch); DVDs can correct about 2800 bytes (a 6-mm scratch).

The compact-laser diode "optical stylus" enables the high capacity of optical memories. A laser beam of wavelength λ in optics with numeric aperture NA can be focused to a diffraction-limited spot diameter [5] of $0.6\,\lambda/\mathrm{NA}$ and depth of focus $0.8\,\lambda/\mathrm{NA}^2$. Typical CD semiconductor infrared diodes have $\lambda = 780$ nm, and head $\mathrm{NA} = 0.5$, giving a full-width, half-maximum (FWHM) spot diameter of about 0.9 μm, and a depth of focus of 2.5 μm. Focus, track wobble, and disc runout pose demanding requirements for disc control servos at these dimensions. Table 1 compares existing disc memories.

3. COMPACT DISCS

3.1. Plain CDs

Compact discs remain the most common optical memories. CDs have mastered marks that are about 0.6 μm wide and from 0.83 to 1.7 μm long. Data are encoded in a non-return-to-zero-inverted (NRZI) format, so 1s (ones) occur at both

Table 1. Comparison of Commercial Optical Memories[a]

Type	CD-ROM	DVD-ROM	DVD-RAM	DVD-R	DVD-RW	MO
Capacity (GB)	0.68	4.7–17.1	4.7 or 9.4	4.7 or 9.4	4.7 or 9.4	2.6
Transfer rate (Mbps)	1.23	11.08	22.16	11.08	11.08	31.2
Rewrite Cycles	NA	NA	>100,000	NA	>1000	>1,000,000

[a]Transfer rates are for single speed drives. CD-ROM 40× drives can deliver up to Mbps transfer rate under constant angular velocity readout. DVD capacity ranges depend on layers in the media. Rewrite cycles are ultimately limited by thermally induced degradation of the active media layer. GB = 10^9 bytes.

edges of marks and 0s are clocked within marks and on the intervening lands. Marks, called "pits" because they are about 130 nm below the surface of the surrounding land, are approximately $\lambda/4$ deep for a 780-nm read laser in polycarbonate ($n = 1.58$). A diffraction pattern arises when the read beam overlaps a pit and its adjacent land regions (Fig. 4). Some diffracted modes do not reenter the read lens, so reading over a pit returns less light than a flat land area.

CD digital audio established the basis for the CD product family. CD-DA can hold up to 74 min of audio, played at 150 kbps, and organized in up to 99 "tracks" per disc. One continuous spiral of data is written. A single directory area after leadin locates tracks on the spiral. Multiple recording sessions or incremental writes are not possible.

The *Red Book* standard (ISO 10149) specifies CD formats [6,7]. Both level 0 (physical layer) and level 1 (sector and track formatting) specifications are given. Level 0 requires cross-interleaved Reed–Solomon code (CIRC) for error correction and EFM for establishing a (2,10) run-length-limited (RLL) disc format. RLL encoding improves mark edge detection, synchronization, and frequency management in playback.

3.2. CD-ROMs

The *Yellow Book* standard established CD-ROM. The CD-ROM sector provides 2048 data bytes (compatible with computer data structures), and uses the rest of the standard sector for additional error correction (mode 1 CD-ROM). The Yellow Book standard also specifies logical sector and logical file organization, and allows retention of a Red Book audio region with CIRC error correction on the disc (mode 2 CD-ROM). Mode 2 access times were slow, as players moved between computer applications (video clips) and the accompanying audio, in later sectors, making multimedia applications unsatisfactory. Mode 1 Yellow Book CD-ROMs can store computer data, compressed audio, and compressed video, and have stronger error correction than mode 2. A Yellow Book extension for multimedia, CD-ROM XA (extended architecture), stores compressed audio (adaptive differential PCM, ADPCM) close to the multimedia sectors to allow smooth access to images and sound. This also allows efficient compressed audio storage. Up to 18 h of monaural sound can be stored on one CD-ROM XA.

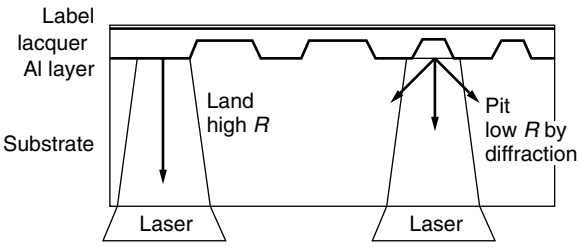


Figure 4. CD disc diffraction structure. As the laser read beam focuses on the flat land, most light is reflected back to the detector by the aluminum layer. Pits and their adjacent off-track land diffract the laser beam away from the objective, yielding a lower effective reflectivity. Polycarbonate substrates are 1.2 mm thick, while pits are only 130 nm deep, and lie just below the top surface.

Specialized standards evolved as CD-ROM was expanded into more demanding applications. The *Green Book* standard addressed efficient storage and access of a mix of data types, primarily for set-top interactive devices, such as Phillips Compact Disc Interactive (CD-I). The same track holds different data types, with each sector having a field that identifies its data type. For example, four levels of sound, from Red Book "CD quality" to various ADPCM levels of quality can be mixed with video, MPEG 1 and 2, and still pictures. CD-I has not made much market impact, and DVD is making the Green Book standard obsolete. The *White Book* standard addresses video CD (CD-V), CD-I, and PhotoCD exchange. While CD-V has been successful in China, Japan, and parts of Europe, it has had little impact in North America and is being replaced by DVD. PhotoCD is evolving its own hybrid standard from the Green Book and Orange Book standards.

The ISO 9660 standard, which evolved from the High Sierra File (HSF) format, addresses file structure and naming problems arising in cross-platform CD-ROM compatibility. ISO 9660 provides a common file structure to CD-ROM developers and is now used by most CD-ROM, CD-R, and CD-RW systems to convert the raw sector data to files, directories, and volumes. There continue to be modifications to the ISO 9660 standard, including Universal Disc Format (UDF) [8] for DVDs and UDF Bridge, which gives backward compatibility to ISO 9660 for readers that accept CD-ROM and DVD.

3.3. CD-R Media

CD-R media, initially called *CD write-once* (CD-WO) or *write-once read-many* (WORM) media, support permanent recording of a single CD. CD-R allows inexpensive, removable, archival optical memory, suitable for limited distribution of applications, images, video, audio, or data, testing of CDs before mass replication and mastering of data images for sending to a disc replicator. CD-R media store up to 650 MB of data and are compatible with all more recent drives under ISO 9660. Data must be streamed continuously throughout each session recording, or the directory information will be incorrect or missing. CD-R is not erasable, so errors in recording cannot be corrected. Typically an image of a complete session is created in a hard-drive partition or on another CD-R. Data buffering gives some security against short interruptions during writing.

On the physical layer, CD-R discs have an active dye layer of cyanine (greenish), phthalocyanine (gold-green), or metal azo (blue) between the polycarbonate substrate and the reflective metallic layer. During writing the dyes absorb heat under the laser spot and permanently change their local reflectivity. Dyes are designed to give high reflectivity change, high speed of response, good consistency, and long shelf life. Early ablative WORM recording thermally vaporized pits in a thin surface metallic film, giving us the term "burning" a CD. Some drive manufacturers wobble the mastered tracking groove to produce extended capacity CD-R discs storing about 700 MB, but these may not be compatible on playback

with all drives. Poor quality dyes on low-cost media may also prevent drives from reading CD-R discs.

The *Orange Book* standard specifies how the various CD application standards should be written onto recordable media, including not only CD-R but also CD-RW and magnetooptic (CD-MO and MO). Orange Book Part I discusses MO systems; Part II describes recording on CD-R, CD-WO and WORM devices; Part III discusses CD-RW. The Orange Book standard defines multisession recording, specifically how data should be stored in the data sectors, where track, session, and directory data are located, how disc and session lead-in and leadout are handled, and where write laser calibration regions occur. Each session of recording requires about 13 MB of overhead for its leadin/leadout and directory areas.

3.4. CD-RW

Rewritable CD-ROM, CD-RW, is the most versatile form of CD optical memory, allowing archiving, testing, distribution, and mastering for replication like the CD-R, with the addition of erasability. The physical data recording process, and the organization of the disc differ from CD-R, and drive requirements are more stringent.

CD-RW media have a polycarbonate substrate, with a mastered pregrooved spiral track. An optical stack manages heat absorption and diffusion during writing and facilitates reflectivity-based readout. The active layer is a metallic semiconductor, usually GST ($Ge_2Sb_{2.3}Te_5$), or AIST (AgInSbTe). These switch from a crystalline state to an amorphous state when the film is heated above its melt temperature. A weaker laser pulse, reaching only the glass transition temperature, returns an amorphous region to crystalline. Since these states have different reflectivity, marks can be written. CD-RW can perform direct overwrite (DOW), allowing faster data transfers. Phase change reflectivity signals are weaker than CD-ROM or CD-R signals, so older drives cannot read CD-RW media. Newer multiread drives have automatic gain control to adjust laser power to CD-RW signals.

In addition to Orange Book requirements, the Universal Disc Format packet writing scheme can be used to give CD-RW compatibility with DVD players. Under UDF, CD-RW discs must be formatted with logical sectors, a process that takes up to 30 min. Preformatted CD-RW media are entering the market. Depending on formatting, a CD-RW contains from 650 MB to 535 MB of user data. CD-RW phase change materials suffer from limited cycle lifetimes. CD-RW media must support at least 1000 write–read cycles; some media manufacturers claim 10,000 cycles.

4. DVD

DVD-ROM is the baseline digital data storage standard for all DVD applications, including DVD video and DVD audio [9]. Unlike CDs, where computer data standards evolved from the audio Red Book, DVDs were designed from the beginning for data memory use. DVD-ROM media store their data as pits that modulate readout intensity, similar to CDs.

Several engineering improvements increase bit density from about 480 Mb/in.^{2} (megabits per square inch) on CDs to over 2.2 Gb/in.^{2} on single-layer DVDs. Marks can be made smaller and closer because a higher NA (0.6), and a red (635-nm) laser yield a smaller diffraction-limited spot ($\approx$0.6 μm). The DVD spiraling data pattern is also more compact, with track pitch 0.74 μm and minimum mark spacing 0.4 μm. A DVD-5, which is a one-sided, one-layer DVD, holds approximately 7 times the capacity of a CD-ROM.

The second innovation of the DVD family is multilayered memory. DVD-ROMs are made from two 0.6-mm injection-molded polycarbonate discs, which are bonded together. Individual single-layer discs are mastered much like CDs. A single-sided disc has a data-bearing disc bonded with a spacer disc. For two-layer storage, two one-sided discs can be bonded, or two layers can be fabricated on one disc, which is then bonded with a spacer disc. In a single-sided two-layer disc, the first layer is injection-molded and covered with a semitransparent reflective layer. A second layer is bonded above the reflective layer, stamped with its own data, and coated with a full reflection layer. Two such 0.6-mm discs can be bonded to make a four-layer DVD (Fig. 5). During reading and writing the laser must focus on the appropriate layer. On a two-layer structure, the laser must read through the top layer of data to the deeper second layer, an additional distance of 55 μm. Two-sided discs must be flipped over or have two heads.

DVDs are defined by a set of application standards, also called "books," maintained by an industry consortium. The standards have not had the wide formal review of a standards organization like ISO/IEC, but nonetheless facilitate the spread of compatible media and players. The entire DVD family is based on the DVD-ROM (Book A), which supports computer applications. DVD-Video (Book B), DVD-Audio (Book C), DVD recordable (DVD-R, Book D), and DVD-rewritable (DVD-RAM, Book E) build on the Book A standard. The consortium (Hitachi, Ltd., Matsushita Electric Industrial Co., Ltd., Mitsubishi Electric Corporation, Phillips Electronics N.V., Pioneer Electronics Corporation, Sony Corporation, Thomson Multimedia, Time-Warner Inc., Toshiba Corporation, Victor Company of Japan, Ltd.) has established a

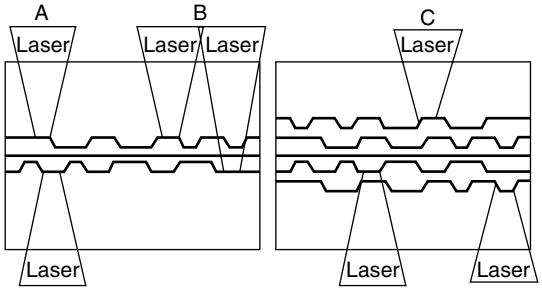


Figure 5. DVD disc structures. Two 0.6 mm CD-like discs are bonded together. A one-sided DVD (not shown) bonds a data disc to a blank substrate. Two-layered DVDs usually bond two data discs, and must be read from two sides (A). Careful reflectivity control allows two-layered DVDs that are read from one side only (B). If two, two-layered discs are bonded together, a four-layered DVD results. These require precise reflectivity control and two-sided reading (C).

jointly owned company (DVD Format/Logo Licensing Corporation) to license DVD formats and logos to third parties and to distribute the DVD Format Books [10].

DVD data are always organized in sectors of 2064 bytes, of which 2048 are data and the rest overhead for addressing, error correction, and copy protection. A combination of 16 byte Reed–Solomon code for the data columns and a 10 byte inner Reed–Solomon code for the rows are used. An $\frac{8}{16}$ run-length modulation is used to generate an acceptable physical sector (4836 bytes), and NRZI transitions produce the actual pit pattern on the disc.

4.1. DVD ROM

The simplest DVD-ROM is the DVD-5, which has a single-layer mastered disc and a blank spacer disc. It is read from one side, at one depth, and has the usual lacquer covering and printed label. Capacity is 4.7 GB per disc. A DVD-9 has two single-layer discs, but is read from one side only. DVD-9 capacity is 8.5 GB. DVD-10 has two single-side discs, each with fully reflective coatings. DVD-10 is read from both sides, and holds 9.4 GB. DVD-18 has two double-layer 0.6-mm discs, each with both partial and full reflection coating, and is read from both sides. Capacity is 17.1 GB. Raw bit densities for all DVD formats remain at about 2.2 Gb/in.^{2}.

A proliferation of DVD applications has created problems with compatibility among DVD memories. DVDs most often use the UDF file system and ISO 9660 file system. As data storage media, they could use other file systems, but this would increase the possibility of drive–media incompatibilities. The DVD Forum has created many application formats, and not all players support all formats. The MultiRead and MultiRead 2 specifications of the Optical Storage Technology Association, and the Multi logo of the DVD Format/Logo Licensing Corporation guarantee that drives will read certain classes of DVDs (and CDs).

4.2. DVD Application Standards

Application standards build on the physical layer specifications of DVD-ROM, file structures such as UDF and ISO 9660, and their extensions. DVD video is an application standard that delivers video, high-performance audio, presentation control information (PCI), and data search information (DSI). DVD video is supported worldwide by computer makers, movie studios, and content publishers. Typical data rates are about 4.7 Mbps, while peak application data rates are over 10 Mbps.

The basic DVD-ROM can accept a variety of video and audio streams, including MPEG-1 and MPEG-2 video, and MPEG-1, MPEG-2, PCM, and Dolby Digital audio, with playing times from one to 13 hours for DVD-5. DVD video media and players should conform to the UDF standard or to the MicroUDF format, which places additional constraints on file structures to simplify consumer electronics.

DVD video supports MPEG-2 video, a choice of multichannel audio formats, and extensive supplemental materials. Players output analog NTSC and PAL, digital interfaces for S-video and HDTV, and different video aspect ratios and display formats.

DVD audio shares many features with DVD video, including storing video and still pictures to accompany audio tracks. This has spawned multiple players, including DVD audio/video, video-capable audio players, and audio-only players. Audio is stored with linear PCM at 16, 20, or 24 bits/sample, with sampling frequency ranging from 44.1 to 192 kHz. Lossy and lossless compression are allowed in the specifications.

The market for DVD-ROM is led by DVD-5, with about 78% of disc releases. DVD-9 and DVD-10 each have about 10% of the total, while DVD-18 has been well under 1%. Almost all these releases have been video titles or games. Player sales rapidly in 2001, and may soon surpass VCR sales.

4.3. Recordable DVD Media

Recordable DVD-R media and players with 4.37-GB capacity began appearing widely in late 2000. The Book E standard is split into DVD-R(A) for authoring and DVD-R(G) for general or home use. These differ in laser wavelengths and land prepit addressing schemes. DVD-R(A) is single-sided, while DVD-R(G) is two-sided and must be flipped over to access both sides. Like CD-R, both DVD-R discs record data permanently by modifying a dye layer, and can support both disc-at-once and session recording. Multilayer DVD-R is not available. DVD-R cannot be erased.

Three rewritable versions of DVD are competing: DVD–RW, DVD–RAM, and DVD+RW. All use phase change materials and support 4.37-GB capacity. +RW and –RW can be rewritten about 1000 times before the phase change materials become unreliable, while RAM, which uses random shifts of the starting write position to reduce media stress, are supposed to withstand up to 100,000 rewrites. RAM and +RW use cartridges, while –RW is usually a bare disc.

DVD-RAM (random-access memory) is the closest product to a fully rerecordable optical memory, and has several technical innovations compared to other DVD rewritable systems. Its data transfer rate is 22.16 Mbps, equivalent to an 18× CD, and marks are recorded both in the land and in the premastered grooves. Zoned linear velocity is used to give good access times. A defect management scheme allows control of manufacturing and formatting defects for more reliable recording.

Rewritable media use *content protection for recordable media* (CPRM) to prevent content theft through unauthorized duplication of DVDs, a major concern of video content suppliers. CPRM places a unique 64-bit media ID in the substrate within the burst cutting area, a band just outside the clamping diameter. The media ID is used to encrypt and decrypt the disc data, such that a rerecorded disc, lacking the media ID, will be unplayable. Other security schemes are being pursued.

5. MAGNETOOPTIC DISCS

Magnetooptic discs offer rewritable, removable, high-capacity optical memory [11,12]. MO media have a shelf life projected to be over 100 years, do not require a

cartridge, and have been rewritten over a million times without losing reliability. One product, the Sony MiniDisc, has been moderately successful, but overall MO discs have not had a widespread market impact on optical memories. MO media and players have been significant in niche markets where high capacity, removability, and long archival lifetimes are important. For most applications MO memories face strong competition from improving magnetic drives and now DVD recordable media.

MO media record in an active layer containing a rare-earth transition metal amorphous alloy, such as TbFeCo, whose magnetic spins prefer to align perpendicularly to the film surface. During thermomagnetic writing, a high-power laser spot heats the active layer to about 250 °C. This elevated temperature reduces the film's magnetic coercivity, allowing a bias field of a few hundred oersted to flip the magnetization. On cooling, the reversed domain persists, creating a mark. Track pitch is typically 1.6 μm, and error-correction coding is similar to CD systems. Run-length encoding of data is used to enhance the readability of the magnetic marks. The MO layer is usually in an optical stack to enhance laser coupling to the metal. A reflective and heat-absorbing layer lies under the stack.

An alternate writing scheme uses magnetooptic magnetic field modulation (MO-MFM) to create domains in a continuously heated stripe under the moving laser beam. This can give higher along-track densities than laser power modulation, but is limited by the dynamics of the biasing magnet.

Readout uses the Kerr effect, in which the polarization of a laser read beam is rotated by the magnetic state of the film. Kerr effect rotation is less than one degree between mark and land, so a differential detection system is needed, with two detectors and polarizing elements in the head. This more complex head, and the bias magnet, require careful design if access speeds are to be maintained. Direct overwrite has also been difficult to achieve, although ingenious use of magnetic multilayers and careful control of pulse power and duration have demonstrated direct overwrite.

Preformatted grooves, including synchronization and header marks, are used to guide the writing and reading processes. The grooves are similar to CD and DVD pregrooves, and create a push–pull tracking and focusing signal. MO capacities are similar to DVD-5, but discs are typically 133 mm. The Orange Book standard applies to commercial MO media.

6. ADVANCED DISC MEMORIES

Near-term improvements in disc-based optical memories will require writing smaller marks to increase areal density, and better focusing and tracking to allow more layers. Several technologies are under development and show strong promise of continuing improvement in capacity and data transfer rate.

6.1. Blue-Violet Lasers

Blue laser diodes, at about $\lambda = 400$ nm, allow increased areal densities in phase change and MO media to over 6 Gb/in.^{2} or 15 GB per disc. Systems for mastering blue laser DVDs are available, but reasonably priced and reliable players must wait for improved blue semiconductor laser diodes. Research in this area is represented by Kondo et al. [13], who report a test on a 19.8-GB single-layer disc, mastered with a 351-nm krypton ion laser, and read with a blue-violet 405-nm laser diode (NA 0.70). Partial-response maximum-likelihood encoding and Viterbi decoding were used.

6.2. Solid Immersion Lens Technologies

Near-field optics [14] optically reduce the size of the coupled spot on the media. Solid immersion lens (SIL) heads incorporate a hemispherical lens that is cut or polished to give an effective NA well above 1.0. A conventional objective lens focuses onto the SIL, which couples a subwavelength spot to the surface. The major drawback is that the SIL must fly above the disc at a height less than 40–80 nm, similar to the flying height of a magnetic head. This compromises removability and robustness of the media head system. Areal density of 50 GB per disc and data transfer of 20 Mbps have been claimed; tracking control remains a problem.

6.3. Novel Laser Configurations

New structures for lasers and new laser-media configurations offer the possibility of higher capacities and data transfer rates. Vertical cavity surface-emitting lasers (VCSELs) are inexpensively manufactured in arrays. Present power levels and wavelengths are not impressive for optical memories, but their array structures suggest the possibility of highly parallel reading and writing. Novel apertured lasers are coated at their front facet, and small apertures are then ion-milled to release a near-field beam smaller than the wavelength. Flying close to the disc, these lasers write and read with a subdiffraction limit spot. A laser-media scheme by Aikio and Howe [15] uses the media itself as part of an external cavity to the read laser. As surface reflectivity is modulated by the data, the reflected light returning to the laser cavity varies and modifies the laser output power. Power can be monitored from the rear facet photodetector common on laser diodes.

6.4. Fluorescent Discs

Frequency multiplexing in the same media volume is possible if the read beam stimulates emission at different wavelengths. This concept underlies the fluorescent multilayered disc (FMD). Many thin active layers would be deposited on a single disc (structures with up to 500 layers have been proposed). With adequate focusing control a small band of layers could be excited by the read beam. If these layers fluoresced at different wavelengths, then one single layer could be filtered and read. A continuing AFOSR project has reported successful design of the head and detectors for this system [16,17].

6.5. Advanced Magnetooptic Systems

SIL near-field optics apply to MO systems, and can be combined with magnetic field modulation and superresolution

multilayered magnetic media to attain extremely high densities. The possibility of MO writing up to 100 Gb/in.^{2} has been claimed with these combined methods [18]. Using just magnetic superresolution with blue lasers has yielded 11 Gb/in.^{2} or 15 GB per disc.

7. HOLOGRAPHIC OPTICAL MEMORIES

After over four decades of effort, holographic memories have shown significant progress recently with impressive laboratory capacities and data rates [19,20]. These demonstrations, while not yet yielding commercial products, have benefited from improved enabling technologies [lasers, spatial light modulators, and CCD (charge-coupled device) cameras] and novel storage media. Holographic memories promise capacities of 125 GB per disc, data rates of 1–10 Gbps, and rapid access times. As a volumetric storage method, their capacities will grow as $1/\lambda^3$ when laser wavelengths get shorter, rather than as $1/\lambda^2$ for surface methods. Both stationary solid media and disc-based media are under study.

Holographic memories store information in a thick photosensitive medium as a phase modulation pattern (hologram) created by the interference of two coherent laser beams: a reference beam and the object or data beam (Fig. 6). The interference pattern of the two beams, captured by changes in the absorption, index of refraction or thickness of the photosensitive medium, can later be probed with a replica of the reference beam. An incident reference beam will be diffracted by the phase pattern to recreate the original data beam, traveling forward as that beam had originally done. It also creates a phase-conjugate beam that travels back toward the original data beam source. This allows reuse of the high-quality writing optics during readout.

Writing is not done bit-by-bit, but rather with a 2-D pattern of data, a page image, typically created by a spatial light modulator (SLM) similar to a LCD display.

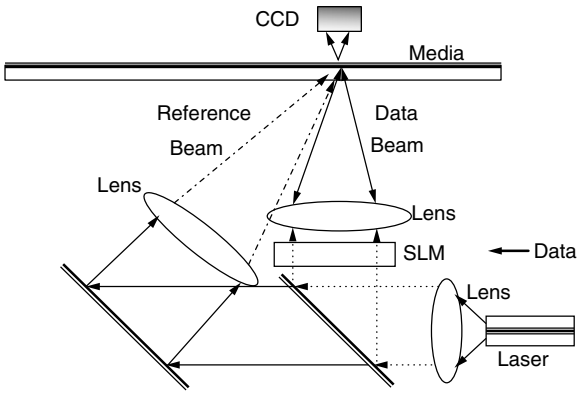


Figure 6. Schematic of a holographic disc system. The laser emits a beam that is collimated and split. One path goes through a spatial light modulator (SLM) that carries a data page, spatially modulating the beam. The second beam is directed as a reference beam. The beams interfere in the disc media creating a hologram indexed by the reference beam angle. On readout the reference beam creates a transmitted beam that is read as a page on the CCD device.

The recreated page image beam can be read by any 2-D array of detectors, e.g., a CCD camera. Multiple pages can be written in the same medium, since individual gratings can be distinguished by multiplexing the incident angle or tuning the laser to another wavelength. Focusing of the beams localizes the holographic patterns in the media. Theoretical capacities depend on the storage media, but exceed 100 Tb/in^3. Considerable technical problems must be overcome to make robust commercial systems based on holography. DARPA has supported the Photo Refractive Information Storage Materials (PRISM) consortium and the Holographic Data Storage System (HDSS) consortium to explore these problems and some progress was demonstrated, including 10 Gb/s data transfer rate [21]. An important tradeoff exists between data transfer rate and media storage density in holographic materials with fixed dynamic range.

Moving a laser probe beam is relatively simple compared to moving a massive read head. Access speeds could therefore be better in holographic memories. Since data is read a page at a time, and since multiple independent probe beams can be read at the same time, aggregate data rates also could be high. Finally, holographic memories offer an associative memory capability that is not possible with any other storage method. When the medium is probed with a search pattern (a *partial* replica of a desired data beam), reconstructed source beams will be created. Each will have intensity proportional to the overlap of their image with the search image. By reading the most intense beam's location with a reference beam, the original data page can be reconstructed.

BIOGRAPHY

Michael F. Ruane received his B.E.E. in 1969 from Villanova University, Villanova, Pennsylvania, and the S.M.E.E. and Ph.D. (Systems) from Massachusetts Institute of Technology, Cambridge, Massachusetts, in 1971 and 1980, respectively. He was a staff member at the MIT Energy Laboratory from 1971 until 1976, and joined Boston University, Boston, Massachusetts in the Electrical and Computer Engineering Department in 1980. At BU he was one of the developers of the Photonics Center, where he maintains the Magnetic and Optical Devices Laboratory, and is also the BU education coordinator for the Center for Subsurface Sensing and Imaging System (CenSSIS). His laboratory studies data storage media and systems, and has contributed to magneto-optical devices, phase change media, disc mastering, and conventional magnetic media. Dr. Ruane has two patents in ellipsometry for optically active media. His areas of interest are in optical systems, the optical storage channel, micromagnetic modeling, and image processing.

BIBLIOGRAPHY

1. H. Varian and P. Lyman, (Oct. 18, 2000) Project Home Page (online), *How Much Information?* Berkeley School of Information Management & Systems, Univ. of California at Berkeley, *http://www.sims.berkeley.edu/research/projects/how-much-info/*, March 30, 2001.

2. C. D. Mee, and E. D. Daniel, eds., *Magnetic Storage Handbook* (*Parts 1 and 2*), 2nd ed., McGraw-Hill, New York, 1996.
3. L. Purcell, *CD-R/DVD: Disc Recording Demystified*, McGraw-Hill, New York, 2000.
4. F. Yu and S. Jutamulia, eds., *Optical Storage and Retrieval*, Marcel Dekker, New York, 1996.
5. Marchant, *Optical Recording: A Technical Overview*, Addison-Wesley, Reading, MA, 1990.
6. International Standards Organization, list of ISO standards related to optical data storage, Search Engine Listing (online), *http://www.iso.ch/cate/3522030.html*, March 30, 2001.
7. D. Chen and J. Neumann, Status of international optical disc standards, *Proc. Recent Advances in Metrology, Characterization, and Standards for Optical Digital Data Discs*, SPIE 3806, 1999.
8. Optical Storage Technology Association, (2001, March 30) reference documents for the Universal Data Format standard and revisions (online), *http://www.osta.org/html/ostaudf.html*, March 30, 2001.
9. J. Taylor, *DVD Demystified*, 2nd ed., McGraw-Hill, New York, 2000.
10. Phillips International N.V., Systems Standards & Licensing, homepage for licensing of DVD technology (online), *http://www.licensing.philips.com*, March 30, 2001.
11. M. Mansuripur, *The Physical Principles of Magneto-optical Recording*, Cambridge Univ. Press, Cambridge, UK, 1995.
12. R. Gambino and T. Suzuki, eds., *Magneto-optical Recording Materials*, IEEE Press, Piscataway, NJ, 2000.
13. T. Kondo et al., 19.8-GB ROM disc readout using a 0.7-NA single objective lens and a violet laser diode, *Proc. Optical Data Storage 2000*, SPIE 4090, 2000, pp. 36–42.
14. T. D. Milster, Near field optics: A new tool for data storage, *Proc. IEEE* **88**(9): 1480–1490 (Sept. 2000).
15. J. Aikio and D. G. Howe, Direct semiconductor laser readout in optical data storage, *Proc. Optical Data Storage 2000*, SPIE 4090, 2000, pp. 56–65.
16. DARPA VLSI Photonics Program Summaries, (2001, January 18). overview page (online), *http://www.darpa.mil/MTO/VLSI/Overviews/Callrecall-4.html*, March 30, 2001.
17. H. Zhang et al., Single-beam two-photon-recorded monolithic multi-layer optical discs, *Proc. Optical Data Storage 2000*, SPIE 4090, 2000, pp. 174–178.
18. D. C. Karns et al., To 100 Gb/in.^{2} and beyond in magnetooptic recording, *Proc. Opt. Data Storage 2000*, SPIE 4090, 2000, pp. 238–245.
19. J. Ashley et al., Holographic data storage, *IBM J. Res. Devel.* **44**: 341–368 (May 2000).
20. H. J. Coufal, D. Psaltis, and G. Sincerbox, eds., *Holographic Data Storage*, Springer-Verlag, Heidelberg, Germany, 2000.
21. National Storage Industry Consortium, description of consortium projects for enhancing magnetic and optical data storage, home page (online), *http://www.nsic.org* March 30, 2001.

FURTHER READING

Annual meetings on optical memory and optical data storage are sponsored by IEEE/Lasers and Electro-Optics Society (LEOS), Optical Society of America (OSA), the International Society for Optical Engineering (SPIE) and other groups. Proceedings appear as SPIE volumes and provide the most convenient access to current research on more advanced optical memories. The most recent volumes include the following:

Mikaelian A. L. ed., *Proc. Optical Memory and Neural Networks*, SPIE 3402, 1998.

Mitkas P. A., and Z. U. Hasan, eds., *Proc. Advanced Optical Memories and Interfaces to Computer Storage*, SPIE 3468, 1998.

Petrov V. V., and S. V. Svechnikov, eds., *Proc. Int. Conf. Optical Storage, Imaging, and Transmission of Information*, SPIE 3055, 1997.

Sincerbox G. T., and J. M. Zavislan, *Selected Papers on Optical Storage*, SPIE Milestone Series, MS-49, 1992.

Sincerbox G. T. *Selected Papers on Holographic Storage*, SPIE Milestone Series, MS-95, 1994.

Proc. Joint Int. Symp. Optical Memory and Optical Data Storage 1999, SPIE 3864, 1999.

OPTICAL MODULATORS — LITHIUM NIOBATE

RANGARAJ MADABHUSHI
Agere Systems, Optical Core Networks Division
Breinigsville, Pennsylvania

1. INTRODUCTION

With the advent of the laser, a great interest in communication, at the optical frequencies, was created. A new era of optical communication was launched in 1970, when an optical fiber, having 20 dB/km attenuation, was fabricated at the Corning Glass Works. Dr. Kaminow and a team from Bell labs reported the concept of electrooptic light modulators [1]. At the same time, Miller [2] coined the term "integrated optics" and heralded the beginning of various efforts in a number of optical components including light sources, waveguide devices, and detectors. The demand for fiberoptic telecommunication systems and larger bandwidth requirements, has increased tremendously since the early 1990s, with the advent of time-division multiplexing (TDM) and wavelength-division multiplexing (WDM) systems. In these systems, the transmitter part basically consists of a laser, which provides the coherent optical (light)wave and the modulator (either external or the direct modulation of lasers), where the desired signal is modulated and is placed on the coherent lightwave.

The direct modulation of lasers is limited by the achievable bandwidth, chirp, or dispersion and the ability to be transmitted to longer distances. The advantages, for short-distance transmission applications include small device size and cost-effectiveness. On the other hand, external modulators are bulky and costly and increase the system requirements. But the advantages, such as large bandwidths and capability to propagate long distances, make these external modulators the winners in optical communication systems. The external modulators include

devices made of dielectric crystals, such as lithium niobate and lithium tantalite; semiconductors such as GaAs, InP, and InGaAs; and polymers such as PMMA. The lithium niobate–based modulators have the advantages of large bandwidth capabilities, low chirp characteristics, low insertion loss, better reliability, and improved manufacturing capabilities. The disadvantages include higher driving voltages, large size of the device, and high cost. The semiconductor modulators have the advantages of smaller size, low driving voltages, relatively low cost (for large volumes), and compatibility of future integration with other semiconductor devices. The disadvantages include large insertion loss, smaller transmission distances, chirp, and manufacturing yields. The polymers are just emerging, and although they can achieve large bandwidths and low driving voltages, the long-term reliability is still being investigated. The $LiNbO_3$ modulator technology, which started in late 1960s, advanced in terms of the material properties, fabrication process, and various modulation schemes in all these years [3–9]. Here, the lithium niobate external modulators are discussed.

1.1. Optical Modulation

It is possible to realize various optical devices, by controlling externally, the lightwave propagating in the optical waveguide. Optical modulators are the devices made of optical waveguides on some material with special properties, where the information is placed on the lightwave externally by imposing time-varying change on the lightwave. The information content is then related to the bandwidth of the imposed variation. Similarly, switches are devices that change the spatial location of the lightwave with respect to the switching signal. These modulators and switches are important components in most of the optical communication systems. The materials may have physical properties, such as electrooptic effect, acoustooptic effect, magnetooptic effect, and thermooptic effect [4].

The modulation types include intensity or amplitude modulation, phase modulation, frequency modulation, and polarization modulation. The intensity modulators are those in which the intensity or amplitude of the coherent lightwave varies according to a time-varying signal. In phase modulation, the phase of the lightwave responds to the applied signal. If the signal is time-varying, the phase change also varies with time. The amplitude of the first sideband and the carrier amplitude are related to the Bessel functions. In polarization modulation, using the electrooptic effect, the polarization states of the lightwave respond to the signal applied. In general, when there is no signal applied, the lightwave emerges as a linearly polarized light. The changes from linear to elliptical polarization, through the applied signal, are characteristics of polarization modulators that use the electrooptic effect. In case of magnetooptic polarization modulators, the light remains linearly polarized but rotated in directions as a function of the applied signal. These polarization modulators are usually used as switches. The last one is frequency modulation, in which the frequency or the wavelength is changed with the applied signal. The detection of such frequency shifts gives rise to more complicated heterodyne system applications.

1.2. Electrooptic Effect

The *electrooptic effect* is, in general, defined as the change of refractive index inside an optical waveguide in optical anisotropic crystals, when an external electric field is applied. If the refractive index changes linearly with the amplitude of the applied field, it is known as the *linear electrooptic effect* or the *Pockels effect*. This effect is the most widely used physical effect for the waveguide modulators. The details can be found in the existing literature [e.g., 4]. Some of the basic fundamentals are given here.

The linear change in the refractive index coefficients due to the applied electric field E_z is given by

$$\Delta n_o = 0.5 r_{13} n_o^3 E_z, \qquad \Delta n_e = 0.5 r_{33} n_e^3 E_z \tag{1}$$

where n_o is the ordinary refractive index, n_e is the extraordinary refractive index, and r_{ij} is the electrooptic constant. For $LiNbO_3$, $r_{33} = 30.8 \times 10^{-12}$ m/V, $r_{13} = 8.6 \times 10^{-12}$ m/V, $r_{22} = 3.4 \times 10^{-12}$ m/V, $r_{33} = 28.0 \times 10^{-12}$ m/V, $n_o = 2.2$, and $n_e = 2.15$ at $\lambda = 1.5$ μm.

2. BASIC STRUCTURE AND CHARACTERISTICS OF THE MODULATORS

In general, the Mach–Zehnder interferometer-type structure is used in the lithium niobate–based intensity modulators. The modulator basically consists of an input divider, an interferometer, and an output combiner. The input divider consists of a straight waveguide and an input Y-branch waveguide, which divides the incoming light into two parts. The interferometer consists of two arms, to which the signal can be applied in the form of voltage. The output combiner consists of an output Y-branch waveguide that combines the two waves from the interferometer arms and finally an output straight waveguide. When there is no signal/voltage applied ($V = 0$), the input wave (field) will be divided into two equal parts, E_A and E_B. At the interference arms they propagate with the same amplitude and phase and recombine at the output Y branch and propagate in the output waveguide without change in intensity (Fig. 1a).

When a voltage is applied, the two waves at the interferometer arms change the phase of the two waves and when the applied voltage, V, is equal to the voltage required, to achieve a π-phase shift, V_π, the output waves from the interferometer have the same amplitude, but a phase difference of π. The output light will become zero by destructive interference (Fig. 1b). For the values of the voltage between V and the V_π the output power varies as

$$P_{\text{out}} = 0.5\{(|E_A| - |E_B|)^2 + 2|E_A| \cdot |E_B| \cos^2 \Delta\varphi\} \tag{2}$$

$$= 0.5 P_{\text{in}} \cdot K_1 + K_2 \cos^2 \frac{\pi V}{2V_\pi} \tag{3}$$

where the phase shift is

$$2\Delta\varphi = \frac{\pi V}{V_\pi} \tag{4}$$

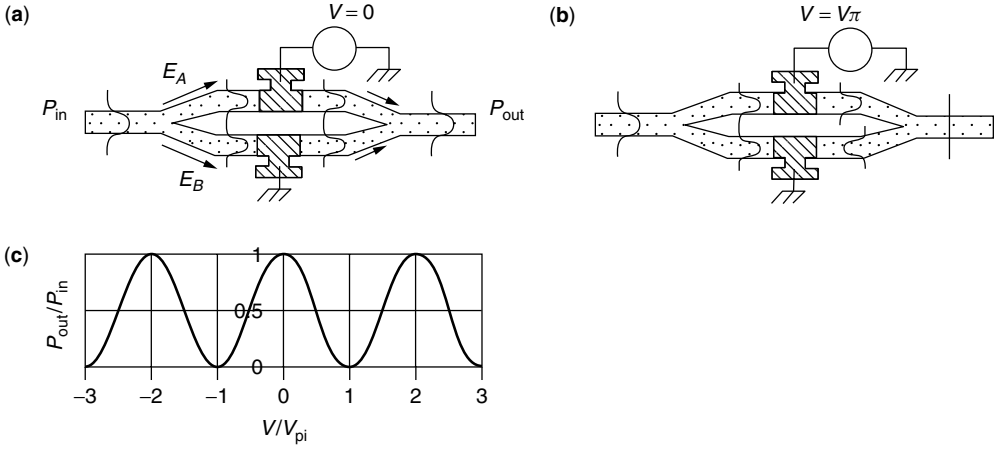


Figure 1. Basic principle of operation of a Mach–Zehnder-type optical modulator, (**a**) without, (**b**) with applied voltage; (**c**) the output intensity as a function of applied voltage.

Figure 1c shows the output intensity as the function of switching/driving voltage, which is represented by Eq. (3).

2.1. Driving Voltage

The change in the index as a function of voltage is

$$\Delta n(V) = \frac{n_e^3 r_{33} V \Gamma}{2G} \tag{5}$$

The phase difference in each arm of the interferometer will be φ, and as the voltage is applied on both arms, the push/pull effect can be used and the total phase difference will be 2φ, where

$$2\varphi = \frac{\pi V}{V_\pi}$$

The voltage length product is

$$V_\pi L = \frac{\lambda G}{2 n_e^3 r_{33} \Gamma} \tag{6}$$

where λ is the wavelength of operation (say, 1.5), n_e is the extraordinary refractive index of the $LiNbO_3$ waveguide (say, 2.15 at λ 1.5 μm), r_{33} is the electrooptic coefficient, 30.8×10^{-12} m/V, V is the voltage applied, Γ is the overlap integral between optical and electric (RF) fields (usually a value of 0.3–0.5), G is the gap between the electrodes, and L is the electrode length.

Depending on the crystal orientation (z-cut, x-cut, or y-cut), the electrodes configuration, whether the electrodes are placed on the waveguides or on the sides of the waveguide, will result in the use of vertical or horizontal fields (Fig. 2).

The overlap integral Γ is better for the z-cut modulator compared to that in the x-cutone. The driving voltage will be less in the case of the z-cut crystal orientation/vertical field, due to the large overlap factor. But there is a need to place a dielectric layer in between the electrode and the waveguides, to minimize the waveguide insertion loss for a TM mode propagation. This will increase the driving voltage. The parameters of the dielectric layer, usually the SiO_2 layer, can be used as a design parameter to achieve larger bandwidths.

2.2. Extinction Ratio and Insertion Loss

If I_o is the intensity at the output of the modulator, when no voltage is applied, $I_{\max}$ is the maximum intensity, and $I_{\min}$ is the minimum intensity when the voltage is applied, then the insertion loss is defined as

$$10 \log \frac{I_{\max}}{I_o} \tag{7}$$

and the extinction ratio (ER) is

$$10 \log \frac{I_{\min}}{I_{\max}}. \tag{8}$$

2.3. Chirp

In case of small-signal applications, the dynamic chirp $\alpha'(t)$ is the instantaneous ratio of the phase modulation to amplitude modulation of the transmitted signal and is expressed as

$$\alpha'(t) = \frac{\dfrac{d\Psi}{dt}}{\left[\left(\dfrac{1}{2I}\right)\left(\dfrac{dI}{dt}\right)\right]} \tag{9}$$

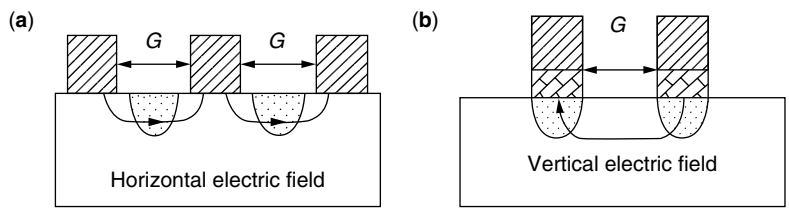


Figure 2. The normally used electrode configurations and the respective field conditions: (**a**) horizontal field used for the x- or y-cut crystal orientations; (**b**) vertical field used for the z-cut crystal orientation.

where Ψ and I are respectively the phase and intensity of the optical field and t denotes the time. In the case of the intensity modulator using the Mach–Zehnder type, the α' can be represented in a simplified form as

$$\alpha' \sim \frac{\Delta\beta_2 + \Delta\beta_1}{\Delta\beta_2 - \Delta\beta_1} = \frac{\Delta V_2 + \Delta V_1}{\Delta V_2 - \Delta V_1} \tag{10}$$

where $\Delta\beta_1$, $\Delta\beta_2$ are the electrooptically induced phase shifts and ΔV_1, ΔV_2 are the peak-to-peak applied voltages of the two arms of the interferometer. Although this expression can be applied in general to the small-signal region, it can also be applicable, to a large extent, for the large-signal region, due to the shape of the switching curve. Also, the value of α' can take minus $(-)$ or plus $(+)$ values, and the chirp can be used to the advantage, depending on the optical transmission system. For systems that operate away from the zero-dispersion wavelength region, and depending on the fiber used for transmission, a negative chirp can be advantageous to achieve low dispersion penalties [10]. In general, for a lithium niobate intensity modulator with a traveling-wave-type electrode, the value can be -0.7. Depending on the crystal orientation and the type of the electrode structure, the value can be zero or can be variable.

3. COMMON ELECTRODE STRUCTURES

A simple electrode structure, consisting of two symmetric electrodes on both interferometer waveguides, otherwise known as "lumped" electrode structure, is shown in Fig. 3a. As the bandwidth, in this case, is limited by the RC (load resistance and modulator capacitance), it is difficult to achieve large bandwidths.

The widely used electrode structure, for large bandwidths, is the traveling-wave electrode structure, where the modulator electrode structure is designed as an extension of the load resistance. Figure 3b shows the structure of a CPW, (coplanar electrode structure), which consists of a central signal electrode and two ground electrodes on both sides of the signal electrode. The two ground electrodes, have widths that are assumed to be sufficiently larger than the signal (or central) electrode structure. Figure 3c shows the asymmetric coplanar stripline (ACPS), or asymmetric stripline (ASL) electrode structure, which consists of a central signal and one ground electrode, where the ground electrode width is assumed to be sufficiently larger than that of the signal electrode. In both of these cases the bandwidth is not limited by the capacitance of the modulator but is dependent on the velocity matching and microwave attenuation of the electrode structures.

The other important characteristics include the following optical characteristics — wavelength of operation, optical return loss, maximum power, and polarization dependency, the following electrooptic and microwave characteristics — bandwidth (frequency response), microwave attenuation, characteristic impedance; and the following mechanical and long term stability — size, temperature, and DC drift stability, humidity, shock, and vibration stability, and fiber pull strength.

These characteristics need to be addressed by the modulator designer, from the initial stage. The waveguide technology is mature enough to satisfy most of the characteristics. The main characteristics that need special attention are the bandwidth and the driving voltage. The usual system requirements are larger bandwidths with lower driving voltages, due to the limitations of available low-driving-voltage drivers. Both bandwidth and driving voltage of lithium niobate modulators are in a tradeoff relationship; one has to be sacrificed for the other. For many years, modulator design has concentrated on optimizing various parameters and finding ways to achieve both larger bandwidths and lower driving voltages [11–19].

The bandwidth of a modulator is dependent on the velocity mismatch between the optical and microwave (RF) and the microwave attenuation of the electrode structure. The velocity mismatch can be controlled by the electrode/buffer-layer parameters. But once the electrode/buffer-layer parameters are fixed, the microwave attenuation (α) is also fixed. In other words, the microwave attenuation, which gets fixed by the electrode/buffer-layer parameters, limits the achievable bandwidth, even though perfect velocity matching is achieved. The driving voltage or V_π is also dependent on the electrode/buffer-layer parameters.

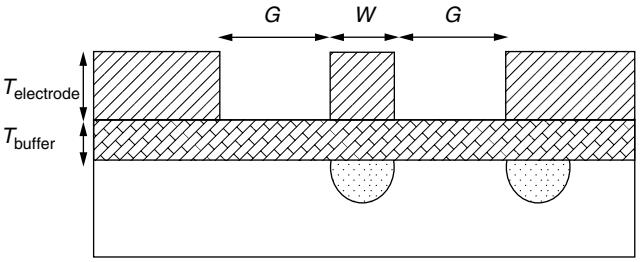


Figure 4. Cross section of a typical Mach–Zehnder optical modulator, with a CPW electrode structure.

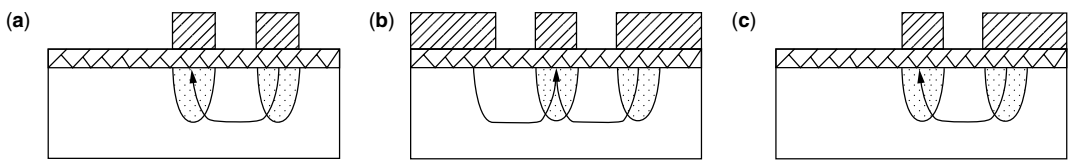


Figure 3. The most commonly used electrode structures: (**a**) lumped; (**b**) coplanar waveguide (CPW) (coplanar electrode structure); (**c**) asymmetric coplanar stripline (ACPS) electrode structure.

As the effective refractive indices of the optical wave (2.15, for TM mode, at 1.55 μm) and that of the microwave (4.2) are different, there exists the velocity mismatch between the two fields, which are propagating simultaneously. This mismatch limits the achievable optical bandwidth value. It is possible to reduce the microwave refractive index to that of the optical refractive index by optimizing the electrode/buffer-layer parameters. Figure 4 shows a cross-sectional view of the Ti-diffused $LiNbO_3$ Mach–Zehnder modulator with a CPW electrode structure. The parameters that are controlled and optimized are W, the width of the signal electrode; G, the gap between the signal and ground electrodes; $T_{\text{electrode}}$, the thickness of the electrode; T_{buffer}, the thickness of the buffer layer; and ε, the dielectric constant of the lithium niobate crystal.

Two-dimensional finite-element analysis can be used for microwave analysis to calculate the capacitance, effective microwave index, and characteristic impedance. The beam propagation method (BPM) or the propagation beam method (PBM) is used for optical field analysis.

The parameters used include the refractive index (TM modes) at 1.55 μm of wavelength, with $n_e = 2.15$, and the dielectric constants of the z-cut $LiNbO_3$ 28 for the z direction and 43 in other directions. The buffer layer is assumed to be SiO_2 with a dielectric constant of 3.9. Figure 5a,b shows the microwave refractive index n_m, and the characteristic impedance, Z, as functions of the electrode width : gap ratio, W/G, buffer-layer thickness, and electrode thickness. It can be observed that n_m decreases with increase in the buffer layer and electrode thickness. These design values depend on various experimental factors and fabrication conditions. Hence, care should be taken in incorporating the experimental values with the modulator design parameters and to ensure that the necessary optimization is performed.

The bandwidth of a modulator can be obtained from the optical response function, which can be defined as

$$H(f) = \frac{[1 - 2e^{-\alpha L}\cos 2u + e^{-2\alpha L}]^{1/2}}{[(\alpha L)^2 + (2u)^2]^{1/2}} \tag{11}$$

where

$$u = \frac{\pi f L(n_m - n_o)}{C} \tag{12}$$

$$\alpha = \frac{\alpha_0 f^{1/2}}{(20\log e)} \tag{13}$$

where α_0 = microwave attenuation constant
f = frequency
n_m = effective microwave index
n_o = effective optical index
$(n_m - n_o)$ = velocity mismatch
L = length of electrode
C = velocity of light

It is evident that even when a perfect velocity matching is achieved, the bandwidth is limited by the microwave attenuation. Thus, reduction of microwave attenuation is the key in achieving very large bandwidths.

The velocity matching using the thick electrodes and thick buffer layer is in the ACPS electrode structure reported by Seino et al. [12]. For an electrode length of 2 cm, a driving voltage of 5.4 V, a bandwidth of 20 GHz, and a microwave attenuation of 0.67 dB/[cm $(\text{GHz})^{1/2}$] were achieved. One problem of the ACPS structure is the resonance problem at higher frequencies, so there is a need to reduce the chip thickness and width. For CPW electrode structure, thick electrodes and buffer layer are utilized [13,14]. For an electrode length of 2.5 cm, a driving voltage of 5 V, and a bandwidth of 20 GHz with a microwave attenuation 0.54 dB/[cm $(\text{GHz})^{1/2}$] was achieved. The issue with a CPW electrode was higher microwave loss due to the higher-order mode propagation. Reduction of chip thickness is needed.

Reduction of microwave attenuation is the main factor in achieving very large bandwidths. The total microwave attenuation of the electrode structure can be reduced by reducing the stripline loss, higher-order mode propagation loss, losses due to bends/tapers, connector, connector-to-pad contact loss, and other package-related loss [20]. A potential reduction of the stripline electrode structure is the use of a two-stage electrode structure. For an electrode length of 4 cm, the driving voltage is 3.3 V, and the bandwidth is 26 GHz with a microwave attenuation of 0.3 dB/[cm $(\text{GHz})^{1/2}$] [21].

3.1. Driving Voltage Reduction

The driving voltage is given by Eq. (6). The driving voltage reduction can be realized mainly by increasing the electrode length or increasing Γ, the overlap integral, between the optical and RF waves, or decreasing G, the gap between the two arms of the interferometer. There is a limit to decrease of G. If the arms are too close, there is a problem of mode coupling between these two arms. This will cause a degradation of the extinction ratio. Also, G is

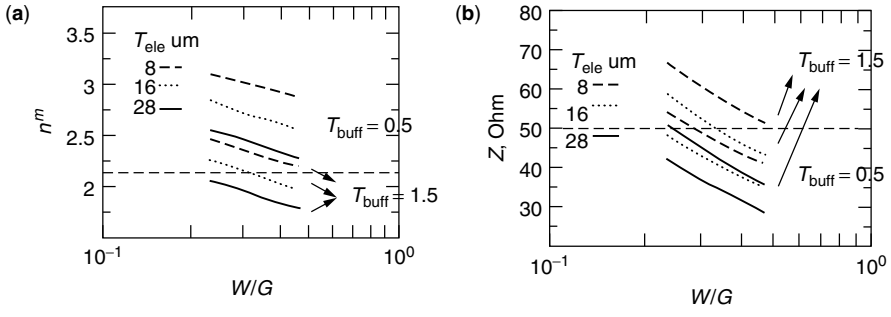


Figure 5. The calculated values of (**a**) microwave refractive index n_m, and (**b**) characteristic impedance, Z, as functions of the electrode width : gap ratio, W/G, buffer layer thickness, and electrode thickness.

the parameter that was fixed in earlier velocity matching design. Increasing the electrode length poses problems on the achievable bandwidth due to microwave attenuation problems. The driving voltage is dependent on the overlap integral between the optical and the microwave fields. The overlap integral needs to be as large as possible, and it depends on the waveguide fabrication parameters and diffusion parameters. The waveguide parameters include the titanium thickness, the titanium concentration, and the gap between the electrodes; the diffusion parameters include the diffusion time and temperature. All these parameters are to be optimized, in order to achieve strong mode confinement. Also, the position of electrodes vis-à-vis the waveguide position dictates the overlap integral value. The other important parameter is buffer-layer thickness, which increases with increase in driving voltage but decrease in overlap integral. Thicker buffer layers are needed to achieve the velocity matching, as explained above. Once the velocity matching condition is obtained, the buffer-layer thickness and the achievable driving voltage are fixed. Optimization of the waveguide/electrode parameters to achieve a strong confinement is usually the remaining issue to achieve the lower driving voltages.

Other methods to reduce the driving voltage include a dual-electrode structure, a ridge waveguide structure, and a controlled buffer-layer structure. In a dual-electrode structure [22], where the two arms of the interferometer are driven by two independent signal electrode structures, the driving voltage can be reduced by approximately half. This structure has the advantage of controlling the chirp value. By individually controlling the voltages applied to the two arms, it is possible to obtain a zero chirp or a negative/positive chirp. In the ridge waveguide structure, by etching ridges in the region, the overlap integral can be increased. At the same time, it is possible to design a modulator to achieve both the velocity matching and the required characteristic impedance. In the controlled buffer-layer structure, the thickness of the buffer layer across the waveguides to achieve both large bandwidth and low driving voltage has been reported [21,23]. The thickness is varied so that both the velocity matching condition and the low driving voltage are achieved at the same time. For the electrode lengths of 4 and 3 cm, driving voltages of 2.5 and 3.3 V, and bandwidths of 25 and 32 GHz were achieved, respectively.

3.2. Reliability

The long-term reliability was the main performance parameter that is vital for using these devices for commercial and practical systems. The DC drift and the temperature stability (and humidity drift) are the main long-term reliability issues [24,25].

3.3. DC Drift

DC drift is the optical output power variation under the constant DC bias voltage application.

Figure 6a shows the output power of the modulator as a function of the applied voltage. The dashed lines show the output power as a function of applied voltage when only AC voltage is applied (and no DC is applied, at $t = 0$), and the solid line shows the same, after $t = t_1$, when DC voltage is also applied in addition to the previous AC signal voltage. The shift between these two curves, ΔV, is the measure of the DC drift. When these types of modulator are used in practical systems, the signal is usually applied at the center of the switching curve (i.e., intermediate between maximum and minimum), which is known as the *driving point*. Once the shift due to DC drift occurs, driving point voltage must be brought back to the previous operating point, using an automatic bias control (ABC) circuit or feedback control (FBC) circuit. It is desirable to minimize this shift, and in most of the cases, a negative shift is more desirable as it facilitates a smaller voltages application through the ABC circuit. The cause of the DC drift can be attributed to the movement of ions, including OH ions, inside the lithium niobate substrate and that inside the buffer layer. It is influenced by the balance of the RC time constants, in both horizontal and vertical directions in the equivalent-circuit model as shown in Fig. 6b. It was also found that the DC drift is affected to a greater extent by the buffer layer. In the circuit model of Fig. 6b, all layers, the $LiNbO_3$ substrate, the Ti : $LiNbO_3$ optical waveguide, and the buffer layer are represented in resistances R, and capacitances C, in both vertical and horizontal directions.

It has been experimentally proved that DC drift can be reduced by decreasing the vertical resistivity of the buffer layer or by increasing the horizontal resistivity of the buffer layer (or that of the surface layer). The surface layer is the boundary layer between the buffer layer and the substrate. The reduction of the vertical resistivity is obtained by doping the SiO_2 buffer layer using TiO_2 and In_2O_3. The increase of the horizontal surface resistivity can be obtained by making a slit in the Si/SiO_2. In both

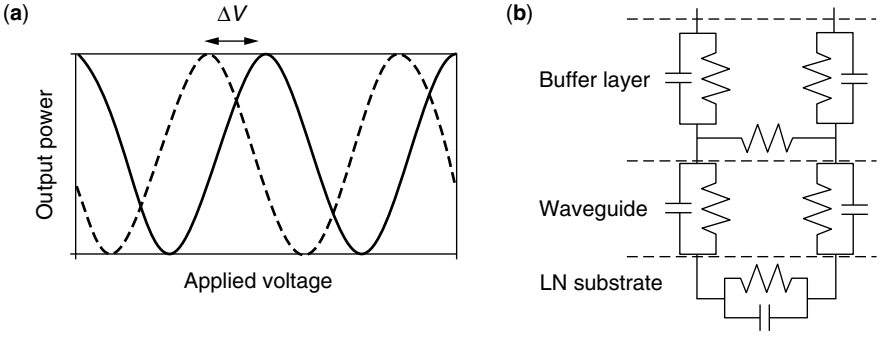


Figure 6. DC drift: (**a**) output power of the modulator as the function of the driving voltage, with and without the DC applied voltage; (**b**) equivalent RC circuit model of the structure, with vertical and horizontal components.

cases, the movement of ions, especially between the two interferometer arms (waveguides), is arrested.

3.4. Thermal Drift

Thermal drift is the optical output power variations as a function of changes in temperature. Once the temperature changes, the piezoelectric charges are induced on the surface of the LN substrate. This causes a surface charge distribution across the two arms of the interferometer, affecting the electric field. This results as a shift in the switching curve and the driving point/operation point shifts, similar to those in DC drift. Hence, in order to reduce this thermal drift, there is a need to distribute or dissipate the charges that are accumulated on the LN surface and between the electrodes. A method in which the charges are dissipated using the semiconductor layers such as Si [26] and other materials was proved to reduce the thermal drift. Another method, in which a Si double slit and a reduction of the resistivity by one order of magnitude, was reported [27].

BIOGRAPHY

Rangaraj Madabhushi received his B.S. and M.S. degrees in physics/mathematics and applied physics in 1974 and 1977, respectively, from Andhra University, Visakhapatnam, India, and a Doctor of Engineering in electronics in 1989 from Tohoku University, Sendai, Japan (optical waveguide devices based on $LiNbO_3$). After working in India, during 1977–1980 (project associate, Indian Institute of Technology, Madras, India) and 1980–1984 (senior scientific officer, Instruments Research and Development Establishment, Dehradun, India), he came to Japan in 1984 on a Japanese government scholarship. After completing a Japanese language course and a doctorate, he joined the NEC corporation, Kawasaki, Japan, in 1989. He worked at the NEC (Central research labs), in various capacities; researcher, assistant manager, manager, on the research, development, and management of $LiNbO_3$ devices, including switches, filters and high-speed modulators for optical communication. In 1999, he came to the United States and joined Lucent Technologies (now Agere Systems) as the technical manager and subsequently promoted as a director in 2000. Since then, he is managing the $LiNbO_3$ and SiWG product development at Breinigsville, Pennsylvania. Dr. Madabhushi holds over 15 Japanese and 10 U.S. patents in the area of $LiNbO_3$ devices and is the author of more than 40 papers in international conferences and journals. He is the senior member of IEEE/LEOS USA, OSA USA, and IEICE, Japan.

BIBLIOGRAPHY

1. I. P. Kaminow, T. J. Bridges, and E. H. Turner, Electrooptic light modulators, *Appl. Opt.* **5**: 1612–1614 (1966).
2. S. E. Miller, Integrated optics: An introduction, *Bell Syst. Tech. J.* **48**: 2059–2069 (1969).
3. H. F. Taylor and Y. Yariv, Guided wave optics, *Proc. IEEE* **62**: 1044–1060 (1974).
4. T. Tamir, ed., *Integrated Optics*, 2nd ed., Topics in Applied Physics, Springer-Verlag, New York, 1979.
5. R. C. Alferness, Waveguide electrooptic modulators, *IEEE Trans. Microwave Theory Tech.* **MT-30**: 1121–1137 (1982).
6. S. K. Korotky, J. C. Campbell, and H. Nakajima, Special issue on photonic devices and integrated optics, *IEEE J. Quant. Electron.* **QE-27**: 516–849 (1991).
7. K. Komatsu and R. Madabhushi, Gb/s range semiconductor and Ti : $LiNbO_3$ guided-wave optical modulators, *IEICE Trans Electron.* **E79-C**: 3–13 (1996).
8. F. Heismann, S. K. Korotky, and J. J. Veslka, Lithium niobate integrated optics: Selected contemporary devices and system applications, in *Optical Fiber Telecommunications*, Academic Press, New York.
9. R. Madabhushi, *High Speed Modulators for Coding and Encoding*, Short course, SPIE Photonics West, Int. Conf. Jan. 2001.
10. A. H. Gnauck et al., Dispersion penalty reduction using an optical modulator with adjustable chirp, *IEEE Photon. Technol. Lett.* **3**: 916–928 (1991).
11. S. K. Korotky et al., High-speed low-power optical modulator with adjustable chirp parameter, *Proc. Topical Meeting on Integrated Photonics Research*, Monterey, CA, paper TuG2, 1991.
12. M. Seino, N. Mekada, T. Namiki, and H. Nakajima, 33-GHz-cm broadband Ti : $LiNbO_3$ Mach-Zehnder modulator, *Proc. ECOC*, paper ThB22-5, 1989, pp. 433–435.
13. M. Rangaraj, T. Hosoi, and M. Kondo, A wide-band Ti : $LiNbO_3$ optical modulator with a conventional coplanar waveguide type electrode, *IEEE Photon. Technol. Lett.* **4**: 1020–1022 (1992).
14. G. K. Gopalakrishna et al., 40 GHz, low half-voltage Ti : $LiNbO_3$ intensity modulator, *Electron. Lett.* **28**: 826–827 (1992).
15. M. Seino et al., A low DC drift Ti : $LiNbO_3$ modulator assured over 15 years, *Proc. OFC'92*, Post Deadline papers, PD3, 1992.
16. D. W. Dolfi and T. R. Ranganath, 50 GHz velocity matched broad wavelength $LiNbO_3$ modulator with multimode active region, *Electron. Lett.* **28**: 1197–1198 (1992).
17. W. K. Burns, M. M. Hoverton, and R. P. Moeller, Performance and modeling of proton exchanged $LiTaO_3$ branching modulators, *J. Lightwave Technol.* **10**: 1403–1408 (1992).
18. K. Noguchi, O. Mitomi, K. Kawano, and M. Yanagibashi, Highly efficient 40-GHz bandwidth Ti : $LiNbO_3$ optical modulator employing ridge structure, *IEEE Photon. Technol. Lett.* **5**: 52–54 (1993).
19. S. K. Korotky and J. J. Vesalka, RC circuit model of long term Ti : $LiNbO_3$ bias stability, *Technical Digest Topical Meeting on Integrated Photonics Research*, San Francisco, paper FB3, 1994, pp. 187–189.
20. R. Madabhushi and T. Miyakawa, A wide band Ti : $LiNbO_3$ optical modulator with a novel low microwave attenuation CPW electrode structure, *Proc. IOOC'95*, Hong Kong, paper WD1-3, 1995.
21. R. Madabhushi, Y. Uematsu, and M. Kitamura, Wide-band Ti : $LiNbO_3$ optical modulators with reduced microwave attenuation, *Proc. IOOC'97/ECOC'97*, Tu1B, Edinburgh, UK, 1997.
22. S. K. Korotky et al., High-speed low-power optical modulator with adjustable chirp parameter, *Proc. of Topical Meeting on*

Integrated Photonics Research, Monterey, CA, paper TuG2, 1991.

23. R. Madabhushi, Y. Uematsu, K. Fukuchi, and A. Noda, Wideband Ti : $LiNbO_3$ optical modulators for 40 Gb/s applications, *Proc. ECOC'98*, Madrid, Spain, 1998, pp. 547–548.
24. S. Yamada and M. Minakata, DC drift Phenomenon in $LiNbO_3$ optical waveguide devices, *Jpn. J. Appl. Phys.* **20**: 733–737 (1981).
25. M. Seino, T. Nakazawa, M. Doi, and S. Taniguchi, The long term reliability estimation of Ti : $LiNbO_3$ modulator for DC drift, *Proc. IOOC'95*, Hong Kong, paper PD1-8, 1995, pp. 15–16.
26. I. Sawaki, H. Nakajima, M. Seino, and K. Asama, Thermally stabilized z-cut Ti : $LiNbO_3$ waveguide switch, *Proc. CLEO'86*, paper MF2, 1986, pp. 46–47.
27. T. Kambe et al., Highly reliable & high performance Ti : $LiNbO_3$ optical modulators, *Proc. LEOS'98*, Florida (USA), Orlando, paper ThI5, 1998, pp. 87–88.

OPTICAL MULTIPLEXING AND DEMULTIPLEXING

ALEXANDROS STAVDAS
National Technical University of Athens
Athens, Greece

1. INTRODUCTION

Optical multiplexing is a technique used in optical fiber networks for enhancing the capacity of point-to-point links as well as for simplifying the routing process within the optical layer. It is found in two forms. Borrowing the concept from its historic predecessor FDM, the optical domain equivalent [which is termed wavelength-division multiplexing (WDM)] is the predominant type of optical multiplexing for reasons to be explained in the following paragraphs. In WDM several information bearers, that is, optical carrier wavelengths, each modulated by a separate data pattern, are launched into (multiplexed) or decoupled from (demultiplexed) an optical fiber (Fig. 1a). The other technique first used in the PCM systems, namely, TDM, also has its optical equivalent, *optical time-division multiplexing* (OTDM). In OTDM, two (or more) pulses of equal energy from the same carrier wavelength are interleaved in time (Fig. 1b). To upgrade a system with bit rate B (pulse duration T) to a system with bit rate $2B$ (pulse duration $T/2$) using OTDM, the following steps are required: (1) generation of pulses with duration $T/2$ in time (at twice the initial power) and (2) delay, probably passive of one stream for $T/2$ with respect to the other before interleaving.

Wavelength multiplexing as a concept exists since the early days of the optical fiber revolution [1]. However, it emerged as a realistic solution only at the end of the 1980s, thanks to optical amplifiers. The advent of the erbium-doped fiber amplifier (EDFA) not only allowed viewing the optical fiber as a "lossless pipe" but also paved the way for collective power restoration of many

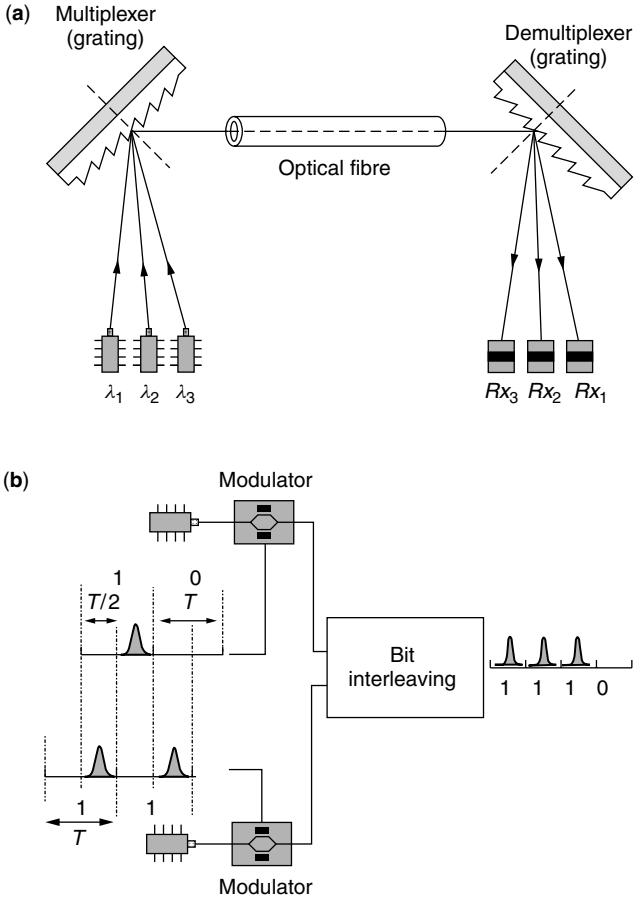


Figure 1. (**a**) Three wavelength channels are multiplexed, transmitted through an optical fiber, demultiplexed at the end and detected from the corresponding receivers (WDM); (**b**) to OTDM two bit streams, two pulses of half duration the initial time are derived (using active components) and a mutual time delay by a half time slot is introduced before interleaving (using passive components).

optical signals simultaneously. For comparison purposes it is mentioned that the first fiberoptic systems employed a single wavelength per fiber were expensive optoelectronic repeaters were used to compensate for the distortions (primarily power attenuation) due to transmission through the optical fiber. Capacity upgrade was achieved by deploying many single-wavelength fibers, something that increased the number of regenerators linearly. In contrast, the EDFA with the ability to restore the power of multiple wavelength channels when they are transmitted over the same single fiber made possible the replacement of these regenerators, leading to enormous cost savings. Wavelength multiplexing was first introduced in the field when it was realized that there is an economic incentive to upgrade from 2.5 to 10 Gbps (gigabits per second) using four wavelengths at 2.5 Gbps instead of one wavelength at 10 Gbps.

In addition, this form of optical multiplexing offers significant routing simplifications within the optical transport layer. Until the 1990s in all commercial telecommunications systems there were only electronic

switching fabrics, rendering electronic switching (and data processing) at the line rate mandatory. Given that the largest fraction of traffic at any node, like in a SDH ring, is transit traffic, processing of the entire traffic volume was becoming progressively more difficult, especially at increasingly higher bit rates. Adopting the principle of wavelength routing, where the final destination is uniquely identified by the wavelength (frequency) of the carrier, it is possible to isolate and process the information content of just the local traffic while the transit traffic, at a different carrier wavelength, could get through the node intact. In current commercial systems with capacity that is scaling up to Tbps (and even tens of Tbps in future systems), wavelength multiplexing is the indispensable technique for capacity upgrade.

On the component level there is a fundamental difference in the nature of the devices used for WDM and for OTDM. In OTDM, the pulses of two or more lower-speed sources are interleaved, generating a datastream with a speed equal to that of the aggregate rate. Thus, very fast "active" (i.e., electrically controlled) devices are needed. For optical processing of this stream in contrast, wavelength multiplexing makes use of "passive" devices. It is the passive nature of these devices that gives WDM all these desirable characteristics generally identified as "transparency": bit-rate independence as well as modulation format and protocol insensitivity. Because of its dominant role in real telecommunication systems, we will consider only wavelength multiplexing for the remaining part of this article.

2. OPTICAL (DE)MULTIPLEXING DEVICES

2.1. Physics of the Devices

When seen from the point of view of technical applications, the most important phenomena of light are *interference* and *diffraction*. Hence, the techniques used for optical (de)multiplexing (regardless of the form in which they appear) are primarily based on one of them. There is no satisfactory explanation of the difference between these two terms [2], but for any practical reason when two optical sources interfere, the result is called *interference*, while when there are a large number of them, the term *diffraction* is more appropriate. For optical (de)multiplexing purposes, the exploitation of two-beam interference is made through devices based on division of the amplitude of the incident beam before they are superposed again. Under this category are devices such as the Mach–Zehnder (MZI), Michelson (MMI), and Sagnac (SI) interferometers. An important family of (de)multiplexing devices are based on arrangements involving multiple divisions of the amplitude or multiple divisions of the wavefront of the incoming wave and they are classified as either (1) interference filters (Fabry–Perot interferometers, multilayer thin-film filters and fiber Bragg gratings) or (2) diffraction gratings (integrated optic, free-space or acoustooptic devices), respectively.

2.2. Functionality

The choice of the technology to be used strongly depends on the type of application under consideration. Hence, for low- to medium-capacity networks, that is, for up to 8-wavelength-channel WDM systems (with bit rates ranging from 644 Mbps to 10 Gbps per wavelength), all the aforementioned devices could be used indistinguishably (Fig. 2). When the total number of wavelength channels N is the predominant consideration for the choice of technology (in particular, when $N \geq 32$), the diffraction gratings are the primary candidates. Nevertheless, regardless of the technological platform, a higher wavelength channel count can be obtained by adding up groups of band-optimized devices. For example, (de)multiplexing devices with up to 60 channels are commercially available using interleaving of band-optimized interference filters in a parallel or cascadable configuration (Fig. 2c,d), while with band-optimized diffraction gratings, several hundred to thousands of channels could be produced.

In Fig. 2, the four different arrangements produce the same final result from a systems point of view. In Fig. 2b, the star coupler facilitates in distributing the same multiwavelength signal to all its N ports (each one will collect $1/N$ of the original optical power). Then a *thin-film* (or a *fiber-grating*) filter will select the requested wavelength. From a functionality point of view, the final outcome is the same as if a diffraction grating is used.* In any case, the diffraction grating-based devices are expected to dominate in the high-capacity systems and, therefore, will be dealt in more detail here. In Section 3 a more detailed comparison between the technological platforms will be provided.

2.3. Diffraction Gratings

2.3.1. Principle of Operation. A diffraction grating is any physical arrangement that is able to alter the phase (optical length) between two of its successive elements by a fixed amount. The impact of this progressive phase alteration becomes evident at the far-field intensity distribution. Consider the case of a plane reflection grating (Fig. 3). The incident plane wavefront PQ first reaches point A, which then becomes a source of secondary wavelets, and hence it advances point B. Finally the incident wavefront reaches B, which then becomes a source of secondary wavelets. These wavelets are exceeding those originating from A at the same time. Hence, the path difference from the corresponding points of the two neighboring grooves (spaced by d), as measured at a distant point of observation, is

$$AD - BC = d(\sin\alpha - \sin\beta) \tag{1}$$

If the incident beam is on the same side of the normal as the diffraction beam, then the sign in Eq. (1) should be replaced with a *plus*. For a more detailed presentation, the reader is referred to classic textbooks such as those by Born and Wolf [3] and Longhurst [4]. In any case it can be shown that the far-field intensity distribution of a planar diffraction grating is the same as that of N rectangular

* Nevertheless the performance in terms of crosstalk, losses [see below] might be different.

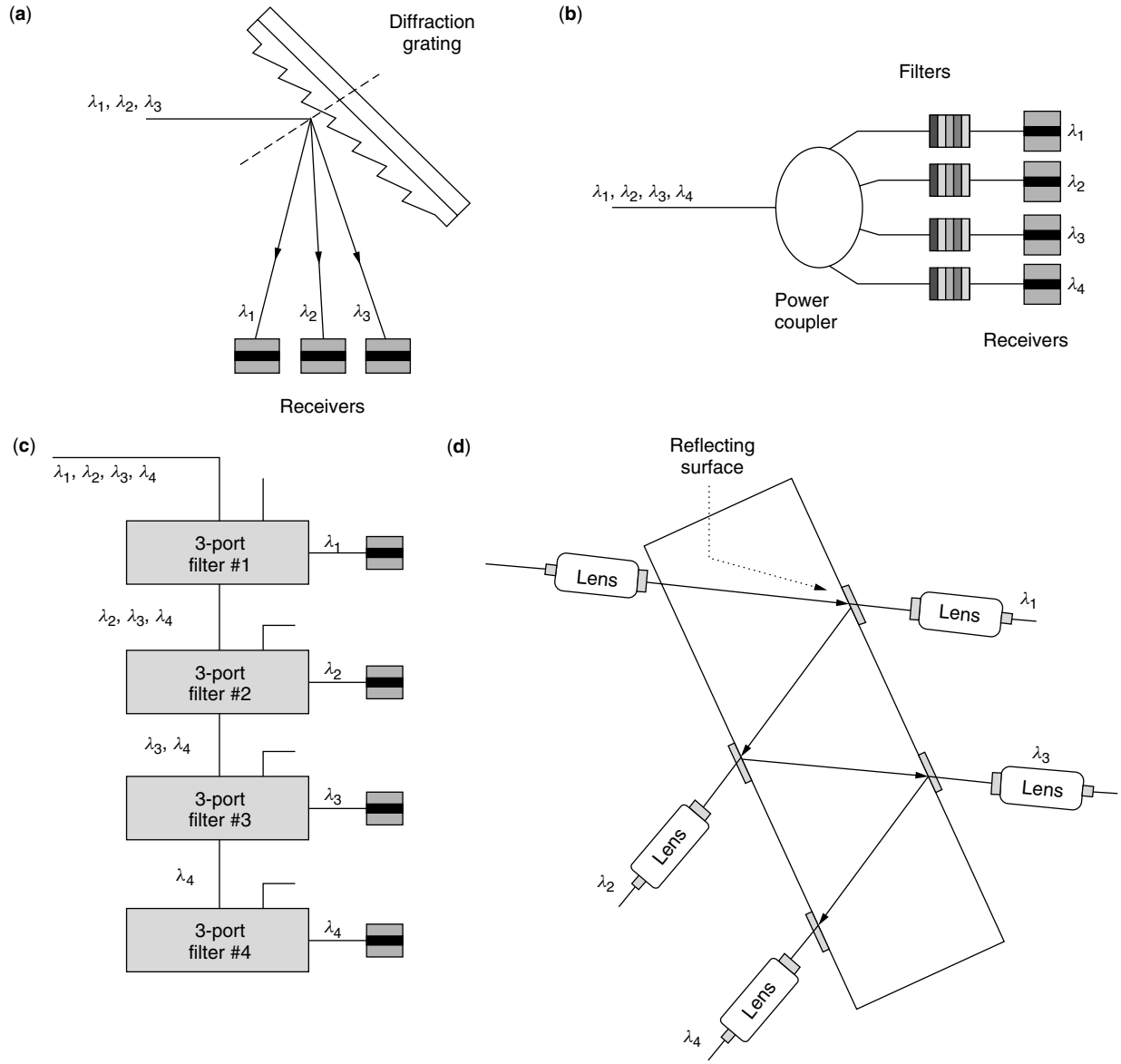


Figure 2. Four equivalent demultiplexing arrangements: (**a**) diffraction grating; (**b**) star coupler (for broadcasting) and fixed-wavelength filters; (**c**) a cascade (bus) of 3-port devices; (**d**) modified "cascade" configuration for two-port devices based on interference (thin-film) filters.

slits modulated by the diffraction envelope of a large slit. Constructive interference occurs when

$$d(\sin\alpha - \sin\beta) = m\lambda \tag{2}$$

where m is a constant called the *diffraction order* and λ is the wavelength (at free space) of the channel. The number N of the grooves/slits determines the sharpness of the principal maximum of the intensity distribution. For example, according to the Rayleigh criterion for the resolution limit, two equal-intensity wavelengths spaced by $\Delta\lambda$ are just resolved if the spatial distance between them is such that the two intensity distributions are crossing each other at 0.8 of their maximum value. The theoretical resolution limit for any diffraction grating is defined as

$$R = \frac{\lambda}{\Delta\lambda} = mN \tag{3}$$

2.3.2. Diffraction Grating Classification. The diffraction gratings could be either concave or planar, and they can operate either at a reflection or transmission mode. A planar diffraction grating — regardless the technological platform and the mode of operation — cannot be used as a standalone component when it is employed as a (de)multiplexer in optical communications. A practical (de)multiplexer based on a planar grating is always implemented in a spectrographic configuration employing two auxiliary optical components: one for collimating the incoming beam (i.e., for transforming the spherical wave to a plane wave) and a telescopic system for focusing the outgoing beam. The wavelength channels are diffracted at different angles determined by Eq. (2), and the telescopic system transforms this angle separation into a spatial separation at the image plane. This spatial

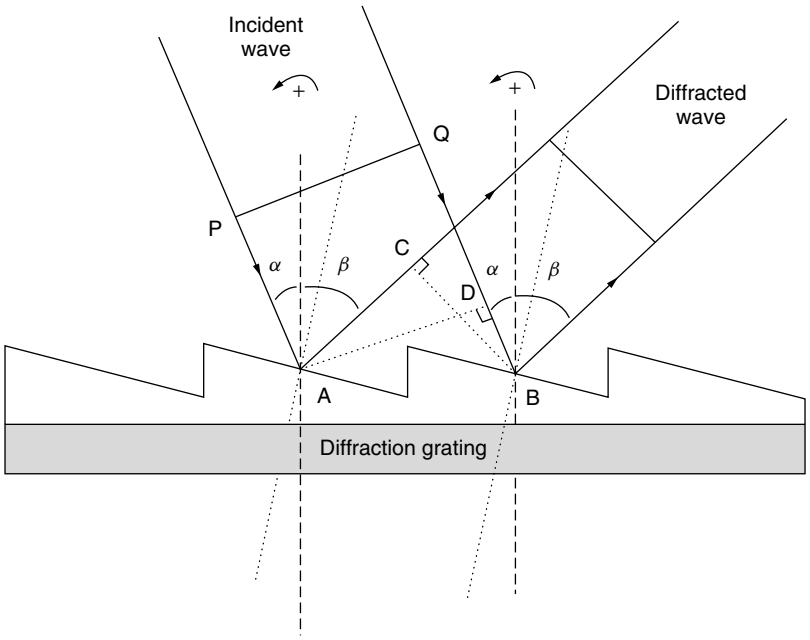


Figure 3. The principle of operation of a planar diffraction grating.

separation (Δx) between wavelength channels ($\Delta\lambda$) in the image plane is given by the reciprocal linear dispersion ($d\lambda/dx$ in nm/mm). Differentiation of Eq. (1) gives

$$\frac{d\lambda}{dx} = \frac{d\lambda}{fd\beta} = \frac{d\cos\beta}{mf} \tag{4}$$

where f is the focal length of the focusing part of the spectroscopic system. In practice, when the spectrograph is used as a (de)multiplexer, Δx is dictated by the minimum distance between the output waveguide/fiber cores.

2.3.3. Spectrograph Overview. The most important configurations for planar grating spectrographs are the *Ebert–Fastie* and the *Czerny–Turner* (Fig. 4a,b). In the former case, the spectrograph is constructed from a planar grating and a large concave mirror (or lens). The *Czerny–Turner* configuration offers the alternative of using two smaller concave mirrors instead of a single large one. The main drawback of these configurations is the use of the auxiliary optics off-axis, something that generates large aberrations. As a result, a point source is imaged as a geometric extended entity that degrades the performance of the optical system (Section 2.4).

A concave grating does not need auxiliary optics since it is a complete spectrograph. The corrugated surface provides the necessary diffraction for wavelength separation or recombination while the geometric properties of the concave surface allow focusing of the diffracted wavelengths. Since concave gratings operate off-axis, they also suffer from large geometric aberrations. However, there are specific geometric arrangements, called *focal curves*, which minimize the adverse effect of these aberrations. The best-known focal curve of concave gratings is the *Rowland circle* (Fig. 4c). For a concave substrate with radius of curvature R, the Rowland circle has a diameter R and is tangent to the apex of the substrate. The important characteristic of this geometric locus is that when a point source A is placed on it ($r_A = OA = R\cos\alpha$), an image free from second- and third- as well as reduced fourth-order Seidel meridional aberrations is produced at a location B on the Rowland circle ($r_B = OB = R\cos\beta$).

2.3.4. (De)multiplexer Performance Considerations. For assessing the quality of any (de)multiplexing device, a number of interrelated issues have to be considered. These include the spectral spacing between adjacent channels, the total number of wavelength channels, the passband flatness, the coupling losses, and the level of outband crosstalk. The best device is the one that allows the largest number of channels with the flattest bandpass, the smallest coupling losses per channel, and the largest optical isolation between adjacent channels with the minimum spectral separation between them.

From Eq. (3) it is concluded that the larger the size of the grating W (i.e., $W = Nd$), the sharper the intensity distribution is and the wider the spatial separation between two wavelength channels can be. Given that sufficiently large optical isolation is available between two adjacent channels, the number of grooves/slits should be considerably greater than that required for satisfying the Rayleigh criterion.

The *coupling loss* of any (de)multiplexer for a given wavelength channel is defined as the ratio of the incoming power to the outgoing power. This could vary with wavelength and depends on many parameters. For diffraction grating devices, these are the propagation material (free-space, Si, III–V semiconductor), the optical aberrations (that depend on the size of auxiliary optics for a planar grating and the clear aperture size for a concave grating), the mode mismatch between the device and the fiber (for integrated optic devices), and the type of the final receptor (e.g., detector, single-mode fiber, or multimode fiber with a clear aperture of 20, 10, and 50 μm, respectively).

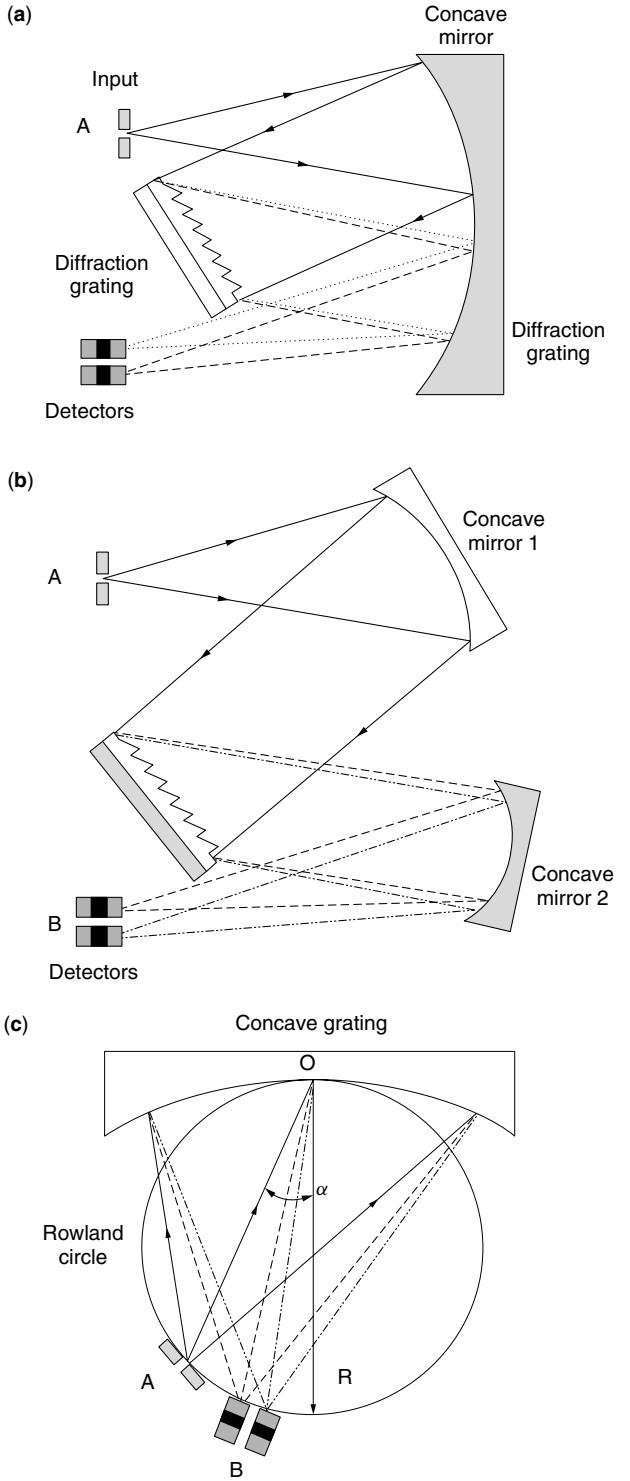


Figure 4. Spectrographs: (**a**) Ebert–Fastie; (**b**) Czerny–Turner; (**c**) Rowland circle.

A fraction of the optical power of a given wavelength that is not coupled to the corresponding outgoing receptor (detector, fiber) could be coupled to the adjacent-channel receptors generating *crosstalk*. In this way, crosstalk is the unintended coupling of signals from adjacent channels due to device imperfections. These phenomena can be better understood by considering the impulse response of any (diffraction grating) demultiplexer (Fig. 5). The impulse response has an intensity distribution with a Gaussian-like central part (due to the Gaussian intensity distribution emitted from a single-mode fiber) and a sinc-square function ($\sin^2 x/x^2$) distribution at the outer parts. As a practical rule of thumb, good optical isolation (low crosstalk) is achieved when the ratio of the optical signal bandwidth (measured at the $1/e^2$ point from its peak value) over the channel spacing is less than 0.25. The fact that the main part of the impulse response has a Gaussian intensity distribution profile leads to passband narrowing when many of these devices are cascaded. For this reason optical techniques are necessary in order to flatten the passband.

2.3.5. Practical Diffraction Grating Devices

2.3.5.1. Arrayed-Waveguide Grating (AWG). The most widely deployed (and studied) type of a grating-based (de)multiplexer is the arrayed-waveguide grating. This is a two-dimensional integrated-optic device (see Refs. 5,6, and references cited within). A special geometric arrangement of two slab waveguides and an array of single-mode waveguides forms a spectrographic setup based on a transmission grating.

The principle of operation, when the device is used as a demultiplexer, is as follows. The multiwavelength channel signal enters the device from the input slab waveguide (Fig. 6a), where it freely propagates. The input and output slab waveguides in most cases are constructed using the Rowland circle (Figs. 3c,6b). In principle, other geometric arrangements (generalized focal curves) are also possible. In any case, aberration-free focal curves are mandatory since the array of the single-mode waveguides is placed at the circumference of a spherical arc. The signal is coupled to the array of the waveguides probably via tapering for best coupling conditions. The array of waveguides plays the role of the grooves/slits in classic gratings [consider Eq. (3)]. Despite the use of the slabs on a Rowland circle, the entire setup has a planar grating configuration (a diffraction grating and two auxiliary optical systems). The length of the array waveguides is chosen such that the optical path length difference ΔL between adjacent waveguides is equal to an integer multiple of the central wavelength of the (de)multiplexer.

Because of the additional phase change introduced by the arrayed-waveguide length difference (that results in "hardwiring" all phases), the corresponding grating [Eq. (2)] is modified to

$$n_s d_0(\sin\alpha + \sin\beta) + n_c \Delta L = m\lambda \tag{5}$$

where n_s is the refractive index of the waveguide slab, n_c is the refractive index of the waveguides in the array (in the most general case, they are not the same), and d_0 is the distance between two successive waveguides in the array. To understand the reasons behind implementation of the AWG using this additional path length difference ΔL, one should consider the following.

Advances in integrated-optic fabrication techniques based on lithographic etching made possible the demonstration of AWGs on InP, Si/SiO_2, or $LiNbO_3$. Despite these

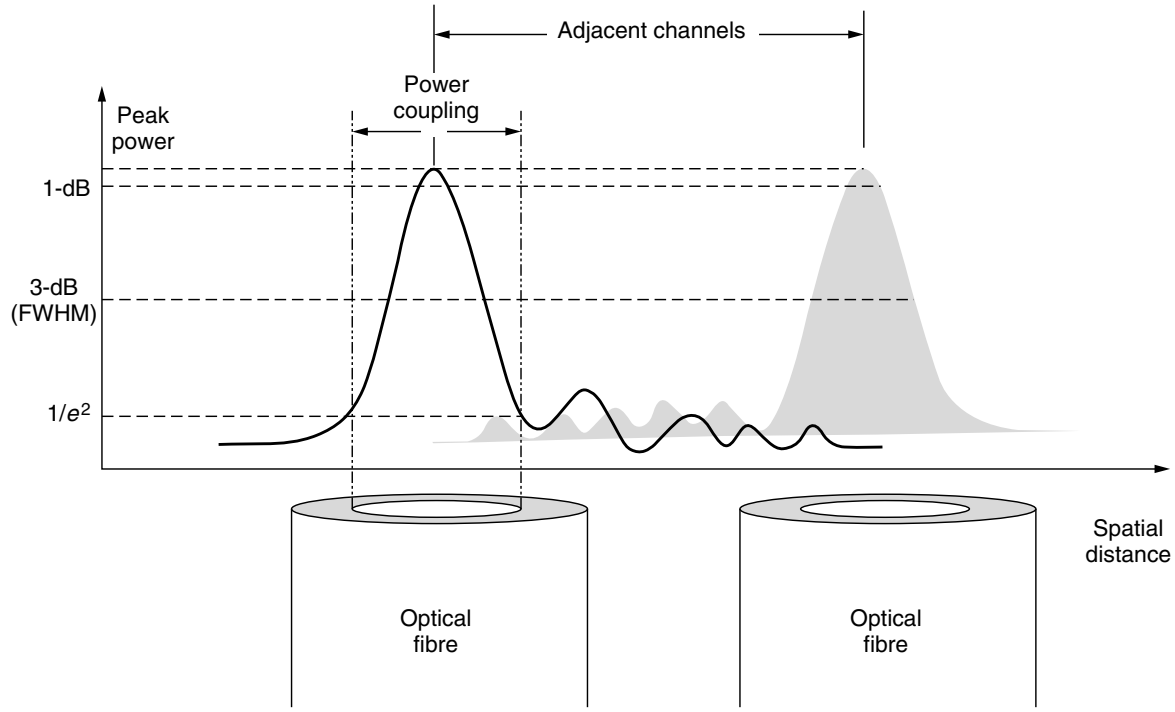


Figure 5. Coupling efficiency and crosstalk between adjacent channels. The part of the intensity not coupled to the destined output fiber smears with the wavelength signal destined to the adjacent fibers.

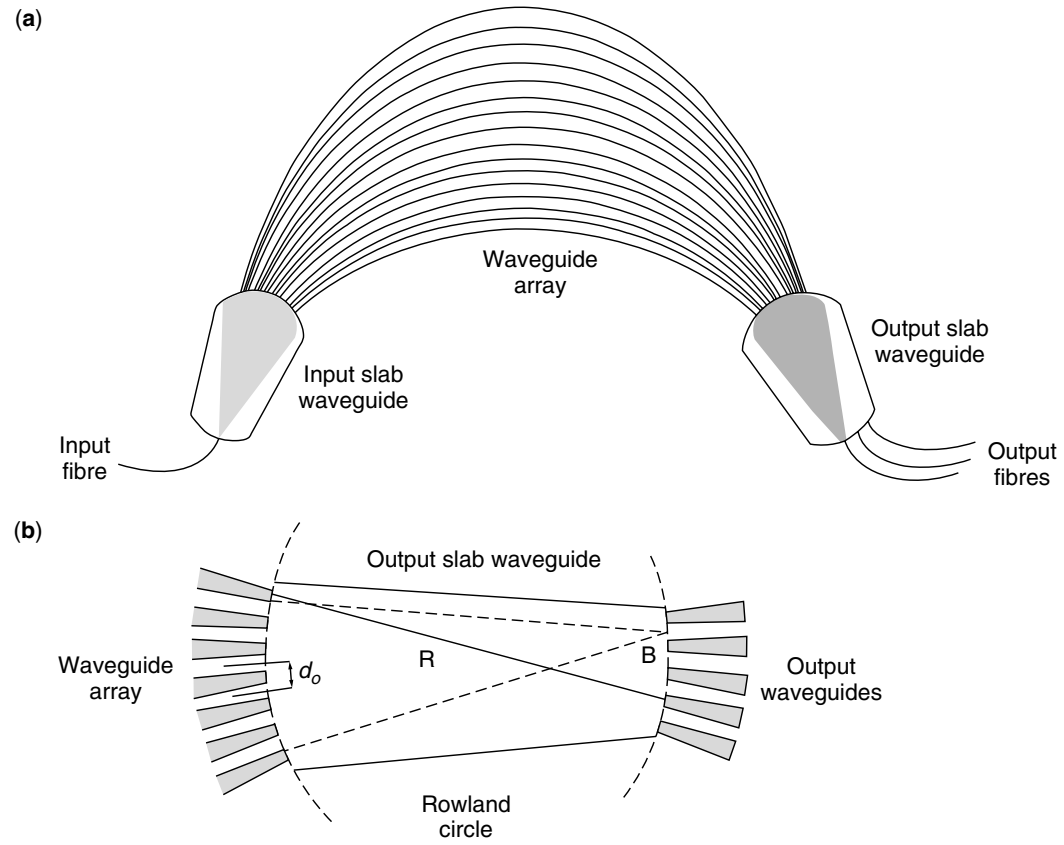


Figure 6. Schematic illustration of the arrayed-waveguide grating (AWG).

advances, there are limitations on the size of the wafers that could be practically used. Because of these size restrictions (resulting in small focal lengths), the required linear dispersion, Eq. (4), is obtained operating the AWG in high diffraction orders. From Eq. (5) the diffraction order equals to $\Delta L = m^{\mathrm{AWG}}\lambda/n_c$ and $m^{\mathrm{AWG}} + 1 = m$. However, the free spectral range (FSR) of the AWG is $\mathrm{FSR} = \Delta\lambda = \lambda/m$, and given that $50 < m < 100$, the FSR < 30 nm. This is one of the main limitations of this approach; specifically, the AWG can be used only in the context of limited spectral range. From an engineering point of view this problem can be solved, as mentioned earlier, by cascading coarse AWG demultiplexers followed by band-optimized fine-granularity AWGs. In this way, a 480-WDM-channel 10-GHz spaced (de)multiplexer has been reported, consisting of a WDM coupler and two 100-GHz-spaced AWGs, followed by 64 10-GHz-spaced AWGs [7]. The largest reported number of channels with a single AWG is 64 channels spaced by 50 GHz (0.4 nm) [8] or 128 channels spaced by 25 GHz [9], both developed on silica.

To improve the performance of the AWGs in terms of diffraction efficiency, modifications of known spectroscopic techniques have been applied [10,11]. The technique requires varying the optical path length (and hence the "hardwired" phase difference) between the waveguides in a nonuniform way, resulting in redistribution of the energy at the image plane. The effect is further improved after a suitable defocusing of the input/output waveguides. The former method is the integrated-optic equivalent of techniques used in aberration-corrected holographic concave gratings where the intensity distribution in the image plane is altered when the corresponding grooves are not equidistant parabolas.

Practical constraints in the fabrication of the AWGs are also attributed to the difficulty of the lithographic system to truly simulate a focal curve such as the Rowland circle. In addition, this mount requires the axis of the input/output waveguides to point toward the pole of the slab, implying that the waveguides have to be tilted with respect to each other. When this condition is not met, the consequent *vignetting* degrades the performance of the outer wavelength channels in the spectrum. Overall, the AWG is a very good candidate device for a demultiplexer operating within a single transmission band (like the *C* band of the EDFA), but it is problematic or impractical for wider optical bandwidth applications. Another issue, common to all integrated-optic devices, is the inherent birefringence of the materials used that leads to TE/TM polarization mode dispersion (PMD)-like problems. Also, temperature controllers are needed to thermally stabilize the operational conditions of the devices.

2.3.5.2. Other Integrated-Optic Spectrographs. Other practical integrated-optic spectrographs used as (de)multiplexers include a *two-dimensional concave grating* [12–15] and a modified *Czerny–Turner* configuration [16]. The former type has been implemented on both silica and III–V semiconductor compounds. Rowland circle or generalized focal curves have been used for producing aberration-free images [13]. The fabrication of this device, in contrast to the AWG, requires deeply etched grating facets, and this is achieved using ion-beam etching. A subsequent problem is the attainable degree of verticality of the grating wall. Another important consideration is associated with the rounding errors of the diffraction grating facets due to lithographic inaccuracies [14]. Again, because of the limited size of the wafers used, the requested linear dispersion [which in the current case is expressed as $d\lambda/dx = n_s d\cos\beta/(mf)$, where n_s is the refractive index of the slab] is achieved by operating the grating at high orders.

The difference between the 2D concave gratings and the AWGs is that in the arrayed-waveguide case the grating constant (pitch) equals to the distance between two successive waveguides. Given that the waveguide length is of the order of a millimeter, the waveguides need to be sufficiently apart to avoid exchange of energy between them (in Si-based devices the distance is at the order of 20 μm). This restriction does not apply to two-dimensional concave gratings. As a result, a grating pitch of few micrometers is feasible and the requested linear dispersion is attained operating the device at a lower order than the AWG (typically at 10th–30th order). Thus, these devices may operate in a wider spectral range compared to their AWG counterparts. With this technology, a device with 120 channels and 0.29-nm channel spacing has been reported [15]. A simple rule for identifying the tradeoff between grating pitch and diffraction order for the integrated optic concave gratings can be obtained by solving the equation for the linear dispersion and the grating equation that leads to

$$\frac{m}{d} = \frac{2\lambda n_s \sin\alpha + 2n_s\sqrt{(d\lambda/dx)^2 f^2 + \lambda^2 - (d\lambda/dx)^2 f^2 \sin^2\alpha}}{2((d\lambda/dx)^2 f^2 + \lambda^2)} \tag{6}$$

On the other hand, the *Czerny–Turner* mount consisting of a transmission grating and two parabolic mirrors used off-axis has been reported [16]. A paraboloid, although the spherical aberration, when it is operated off-axis. Introduces a significant amount of meridional coma. A *Czerny–Turner* spectrograph should be deployed using two spherical mirrors with different radii of curvature in order to compensate for meridional coma that degrade crosstalk.

In any case, both (de)multiplexer types manifest the same dependence for temperature control and compensation for the PMD-like dispersion due to material birefringence as the AWGs.

2.3.5.3. Free-Space Gratings. Practical free-space optical (de)multiplexers can be found in the form of either planar grating mounts or a holographic concave grating. Free-space grating multiplexers were the first to be tested in conjunction with WDM system experiments. The most established planar grating (de)multiplexer is implemented on the basis of a modification of the *Ebert–Fastie* configuration where the source and the image are almost collocated at the optical axis of a parabolic mirror. This is now a commercial product called STIMAX [17]. Operating an optical system on-axis results in an aberration-free image from second- and third-order Seidel aberrations. Fourth-order

Table 1. Performance of Diffraction Grating Devices

	Channel Spacing (nm)	Number of Channels	Losses (dB)	Crosstalk (dB)	Comments
AWG[a]	0.8	40	<6	< −20	Si/SiO_2
AWG[a]	0.8	≤40	<8	<−25 dB	1-dB band, ∼0.16 nm
AWG[a] flat passband	0.8	≤40	<9	<−24 dB	1-dB band, ∼0.32 nm
AWG [9][b]	0.2	128	3.5–5.9	<−16 dB	Si/SiO_2
Free-space planar[a]	0.8	≤40	<5.5	<−30 dB	1-dB band, >28 GHz
STIMAX[a]	0.8	≤64	<5.5	<−30 dB	—
Minilat[a]	0.8	≤92	<8	<−33 dB	1-dB band, ∼0.2 nm
Holographic concave[b] [24]	0.4	64	<8	<−30 dB	—
2D concave[b] [12]	1	50	16	<−19 dB	InP
2D concave[b] [15]	0.29	120	20–40	<−44 dB	Si-based

[a] Commercially available products.
[b] Laboratory results.

aberrations (spherical aberrations) do exist, and they are compensated by means of the parabolic mirror.

The STIMAX (de)multiplexer has a very high wavelength channel count (Table 1). The main limitation of this configuration is the rapid increase of third-order Seidel aberrations (coma) for the outer wavelength channels of the spectrum due to the use of the parabolic mirror off-axis. Mechanical and thermal stability issues have been successfully addressed. As a result of the free-space operation, PMD-like problems are inherently absent while the polarization dependence of the diffraction efficiency is an issue especially for wider bandwidth applications.

Another configuration uses a planar grating together with retroreflectors at the image plane that facilitate in producing a zero-dispersion focused light, reducing bandwidth narrowing due to cascaded (de)multiplexers [18]. In principle, this approach provides a bandwidth flattening technique that is still applicable even when the device is operated as a demultiplexer, something that is not the case with other solutions used in AWGs [19]. A variant of this technique is used to eliminate the polarization dependence of the diffraction efficiency [20].

Concave diffraction gratings are single-element optical systems that simultaneously provide dispersive and focusing properties. A single optical element is very attractive because of easier optical alignment and reduced packaging problems. For this reason concave gratings were used as (de)multiplexers even at the very early stages of WDM transmission [21]. However, it has been recognized that the main disadvantage of this optical system is the large inherent aberrations associated with the spherical concave substrate and, thus, aberration-corrected gratings need to be deployed [22].

The holographic concave gratings are the primary candidates for providing aberration corrected (de)multiplexers [23]. The principle of operation of the holographic concave grating is as follows (Fig. 7). The concave substrate is covered with a suitable photoresist material sensitive to a specific wavelength (e.g., Ar^+ laser). A laser beam is split and focused in two pinholes such that two coherent point sources (C and D) are created. In general, the resulting spherical waves interfere on the substrate, generating fringes that are neither equidistant nor straight. This is the recording phase. When the device is used as a demultiplexer, the single-mode fiber carrying a multiwavelength signal is placed at point A. Then a particular wavelength is imaged at point B, which, however, is not a point source, due to large optical aberrations. The aberrations can be eliminated by placing the two point sources C and D such as the generated fringes introduce a phase shift that cancel out the aberrations introduced by the spherical substrate. It has been demonstrated that up to 1000 wavelength channels can be (de)multiplexed using these devices, covering a spectral range of 200 nm [24]. Nevertheless, these devices are still available only in laboratories. As with all free-space devices, mechanical stability and polarization-dependent diffraction efficiency is an issue, and in particular when the number of wavelength channels are more than a hundred, the problem of an efficient fiber mount has to be tackled. The performance of all diffraction grating devices is summarized in Table 1.

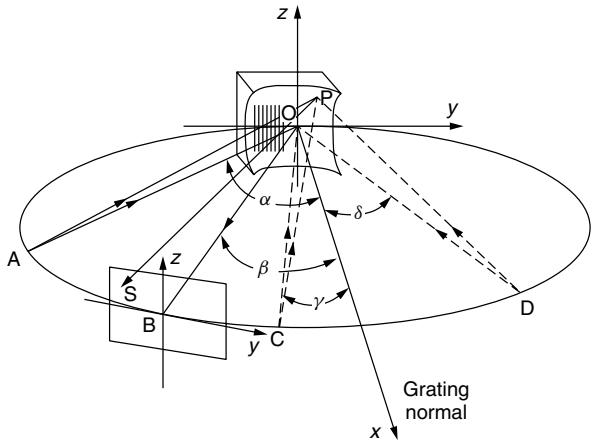


Figure 7. Schematic illustration of a holographic concave grating.

2.3.5.4. Acoustooptic Grating. This is an *active* device. The principle of operation of this device is based on the interaction of light with sound resulting in a transmission

grating. A sinusoidal sound wave travelling at the surface of an appropriate material generates periodic variations of the density (or strain) of the material according to the frequency of the wave. As a result, the macroscopic effect is a periodic change of the refractive index, and these periodic changes act as partially reflecting mirrors. Hence an incident plane wave, when specific conditions are met, will be diffracted at an angle according to its wavelength. The grating formed by the sound wave is a dynamic (time-varying) one.

The effect of the sound wave on the impinging plane wave can be understood in two ways. The distance between two "partially reflecting mirrors" depends on the frequency Ω of the sound wave. Conservation of energy and momentum require that $\omega_i = \omega_d + \Omega$, and $\mathbf{K}_\mathrm{i} = \mathbf{K}_d + \mathbf{K}$, respectively, where the index i indicates the incident wave while the index d the diffracted. $\mathbf{K}$ is the wavevector of the sound wave, and since $\Omega \ll \omega_i$ it is $\omega_i \cong \omega_d$. Thus, apart from a negligible frequency shift, the effect of the sound wave on a multiwavelength signal is to change the direction of propagation according to wavelength (demultiplexing) (Fig. 8a). Alternatively, it could be argued that an optical path length (phase) delay occurs between the two "partially reflecting mirrors" similar to what is produced by the grooves of a diffraction grating. When the distance between them satisfies the grating equation (which now is termed the *Bragg condition*), the elementary planes add constructively. The intensity distribution of the impulse response has the known sinc-square form and its sharpness will depend on the number of reflecting mirrors, namely, the total interaction length L (the sharpness will depend on the ratio L over the wave number determined by Ω).

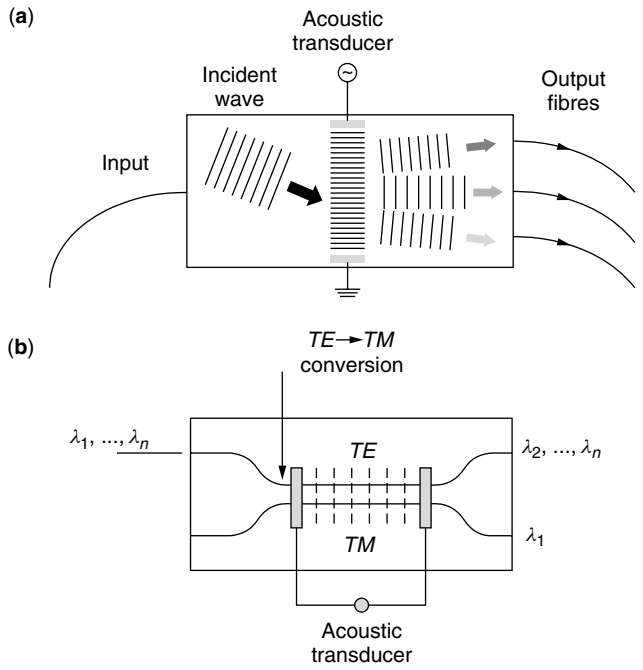


Figure 8. (**a**) The principle of operation of an acoustooptic grating; (**b**) a four-port acoustooptic filter.

2.4. Optical Filters

2.4.1. Acoustooptic Filters (AOFs). When an acoustic wave is applied on an acoustically active material, the induced birefringence (i.e., a dissimilar change of the refractive index for the ordinary and the extraordinary rays) of the medium alters the state of polarization of the incident wave. This principle is used for constructing acoustooptic filters (Fig. 8b). A multiwavelength signal enters a four-port device like a single coupler design. At the input stage, the incoming signal is 3-dB split by the directional coupler and the light at the lower branch also undergoes a phase delay of $\pi/2$. In other words, at the lower branch a polarization rotation is observed from the TE to the TM mode. When no acoustic wave is applied, another $\pi/2$ phase delay occurs at the output part and all channels exiting from the symmetric to the input port (e.g., upper-in, upper-out). When an acoustic wave is applied, the exact matching conditions are altered via the acoustooptic effect and the requested channel is selected from the lower output port. An interesting feature of the acoustooptic filter is that many acoustic frequencies could copropagate, allowing a simultaneous selection of more than one channel. Hence, the AOF can be used as a band-selecting filter in hierarchical (coarse/fine) WDM (de)multiplexing since it has a tunability of hundreds of nanometers. The crosstalk figure of the AOF is not as good as that of the other commercial diffraction gratings.

2.4.2. Interference Filters. These filters appear in the literature under many different names such as dielectric thin films, multilayer interferometric filters, and multistack thin-layer filters. Further, they are constructed from many different compounds ranging from liquid crystals to various oxides (SiO_2 or TiO_2) to multi-quantum-well (MQW) III–V semiconductors. Nevertheless, the principle of operation for all these structures is easily understood considering a Fabry–Perot etalon [25] that is an interferometer based on multiple divisions of the amplitude. Note that collimating optical devices [like bulk or GRID rod lenses] are mandatory at the input/output.

Let us assume that a material A with higher refractive index (n_H) compared to a material B (n_L) forms a cavity as shown in Fig. 9a. The reflected and transmitted intensities I_R and I_T of the Fabry–Perot etalon, respectively, normalized to the incident intensity, are given by

$$\frac{I_R}{I_i} = \frac{4R\sin^2(\delta/2)}{(1-R)^2 + 4R\sin^2(\delta/2)},$$
$$\frac{I_T}{I_i} = \frac{(1-R)^2}{(1-R)^2 + 4R\sin^2(\delta/2)} \tag{7}$$

where $\delta = (4\pi n_H \cos\theta_d L)/\lambda$ and R is the fraction of the intensity reflected at each interface like n_H/n_L and n_L/n_H .

Maximum power transfer to the reflected beam occurs when the thickness of the high-refractivity region is equal to a quarter-wavelength ($\delta/2 = m\pi/2$) while maximum transmission occurs when the thickness is one half-wavelength thick ($\delta/2 = m\pi$), assuming that $R \cong 1$. The former case is depicted in Fig. 9b. Many

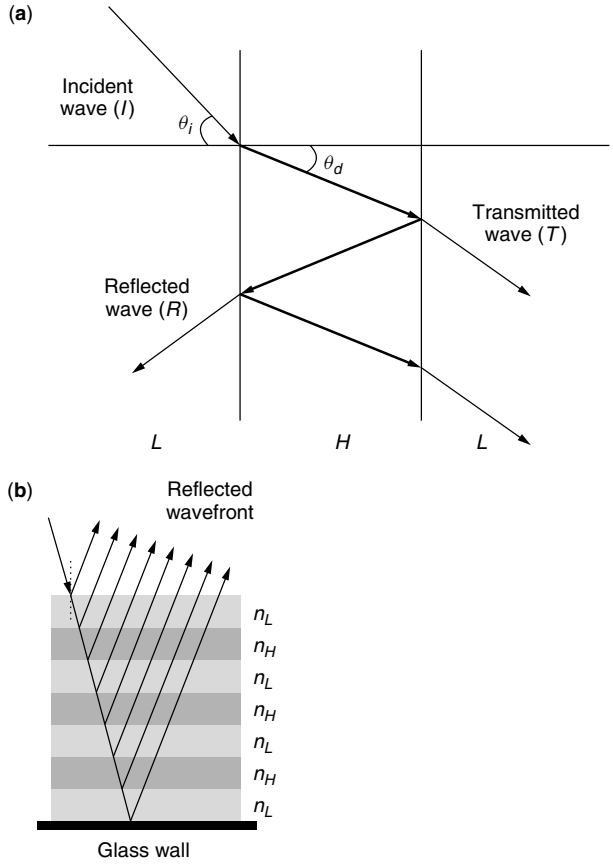


Figure 9. (**a**) A Fabry–Perot etalon; (**b**) a multilayer stack operated on a reflection mode.

different configurations could emerge by adding up such multistacks; for example, two structures as shown in Fig. 9b with a space layer between them form an additional Fabry–Perot cavity of the type HLH(2L)HLH, assuming that each layer has a quarter-wavelength thickness. It can be shown that sharper cutoff characteristics are obtained by increasing the number of layers or cavities. However, when all cavity lengths are the same, the overall structure produces a narrower passband. Hence, layers of unequal thickness are used, something that requires the addition of further layers for phase matching, leading to a complex optimization problem. Also, the passband increases with increasing values of n_H/n_L [26]. Overall, practical constructions lead to filters offering optical isolation up to -30 dB. The losses are range between 1 and 5 dB depending on the material, the number of channels, and other factors.

It is pointed out that these are the only filter devices that could be implemented using all the three configurations of Fig. 3b–d. Nevertheless, the reader should be aware that should the configurations of Fig. 3c,d be used in a practical system, power equalization techniques should be employed for compensating the loss variation (which could be a problem when these devices are cascaded if the number of channels is high).

2.4.3. Mach–Zehnder Filters. These filters are four-port devices that can be either *passive* or *active*. The

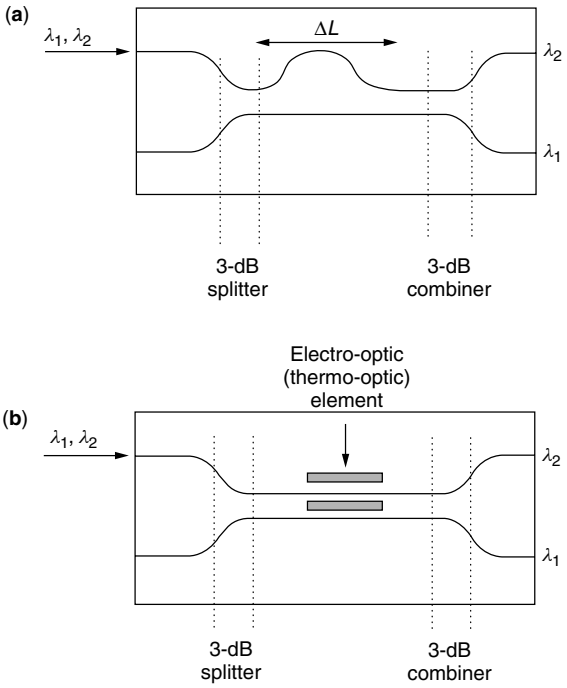


Figure 10. (**a**) The asymmetric Mach–Zehnder and (**b**) the symmetric configurations.

integrated-optic version of the Mach–Zehnder interferometer, illustrated in Fig. 10, consists of three parts. The input and the output parts are 3-dB couplers, while the central section has two waveguide arms (upper and lower) with different path lengths. The configuration is called *asymmetric*. For a signal entering the upper port, the overall phase difference due to the asymmetric length is

$$\Delta\phi = \frac{2\pi n_u}{\lambda}(L + \Delta L) - \frac{2\pi n_l}{\lambda}L = \frac{2\pi n}{\lambda}\Delta L \qquad (8)$$

where n_u and n_l are the refractive indices of the upper and lower waveguides, respectively. In this case they are assumed to be equal to n. The upper ports are labeled as 1 and $1'$ at the input/output, respectively, and likewise the lower ports as 2 and $2'$. The transmittance from port 1 to $1'$ is given by T_u and from 1 to $2'$, by T_l.

$$T_u = \sin^2(\Delta\phi), \quad T_l = \cos^2(\Delta\phi) \qquad (9)$$

when $(2\pi n/\lambda_1)\Delta L = (m + 1)\pi/2$ and $(2\pi n/\lambda_2)\Delta L = m\pi$, with m an integer, then T_u is unity for λ_1 and zero for λ_2 and vice versa (T_l is unity for λ_2 and zero for λ_1). In this way, the wavelengths λ_1 an λ_2 are collected from different ports. When the device is carefully designed, it can operate as a wavelength channel (de)interleaver. The *symmetric* configuration (Fig. 10b) has waveguide arms of equal length, so the phase difference occurs as a result of electrooptically or thermooptically [27] induced change of the refractive index, resulting in different n_u and n_l when the control signal is on.

The loss figure depends on the number of wavelengths that can be demultiplexed from a single module, as well as the host material (Si or III–IV semiconductor). In general,

the optical isolation between adjacent channels is not better than −30 dB.

3. OVERALL ASSESSMENT OF (DE)MULTIPLEXING TECHNIQUES

Having presented the main technological platforms currently used for optical (de)multiplexing, it would be interesting to highlight their pros and cons. As pointed out earlier, the main question when a technology is assessed is the type of application in mind. In general, the configurations illustrated in Fig. 2b,c have different drawbacks.

A layout like the one in Fig. 2b, based on a power coupler and optical filters, is a flexible solution up to approximately 8 wavelength channels. Beyond this point, splitting losses tend to be high that the grating solution is advised. Optical filters with tailormade spectral characteristics (e.g., through wavelength) can be used in this configuration leading to a (de)multiplexer construction, offering a wavelength comb with unequal channel spacing allocation (to combat, e.g., fiber nonlinearities such as four-wave mixing). Nevertheless, when systems with a large number of wavelength channels are desired, the cost of the system scales proportionally to channel count.

The "bus" architecture of Fig. 2c is implemented only via three-port or four-port devices. Thus, the loss performance is not uniform across the spectrum of interest; the first channel has the lowest losses while the final channel suffers from the worst losses. It is this loss figure that determines the maximum number of channels to be used per band. In general, the performance of the optical filters is good in terms of optical isolation. The loss performance of the device itself is good, but for practical applications a cascade of other optical components, such as band (de)multiplexers, is required.

With diffraction gratings the cost is not proportional to channel count and, indeed, devices with a very large number of channels have been demonstrated. AWG suffers from polarization-induced phenomena, while free-space gratings do not. Free-space gratings offer perhaps the best optical isolation from all (de)multiplexing devices and have no restrictions with respect to the total optical bandwidth they can handle. In principle, the free-space devices can explore the parallelism of optics to generate many (de)multiplexers in parallel or to be used in conjunction with other free-space devices such as microelectromechanical switches (MEMSs) in optical cross-connects.

BIOGRAPHY

Alexandros Stavdas received his B.Sc. in physics from the University of Athens, Greece, in 1988, his M.Sc. in optoelectronics and laser devices from Heriot-Watt University/St-Andrews University in 1990, and his Ph.D. from the University College of London, United Kingdom, in 1995 (supervisor, Professor J.E. Midwinter) in the field of wavelength routed WDM networks. He worked in the design of free-space and integrated optics demultiplexers, wavelength cross-connects and on issues related to optical switching and wavelength routing systems. In the past he worked on the ACTS COBNET Project on alternative ring architectures and in design considerations and scalability of WDM rings, and on ACTS PLANET in the area of generic architectures for the WDM upgrade of SuperPONs. Currently, he is leading the project ULTRA funded by Nortel Networks on ultra-wideband DWDM systems and he is the technical leader for NTUA on the IST-DAVID project dealing with packet-over-WDM in Metropolitan Networks. He served as chairman of the Optical Network Design and Modelling Conference (ONDM 2000). Current interests include physical layer modeling of optical networks, ultra-high capacity end-to-end optical networks, OXC architectures, WDM access networks, and optical packet switching.

BIBLIOGRAPHY

1. C. Koester, Wavelength multiplexing in Fiber optics, *J. Opt. Soc. Am.* **58**(1): 63–67 (1968).
2. R. Feynman, *Lectures in Physics*, Addison-Wesley, Reading, MA, 1983.
3. M. Born and E. Wolf, *Principles of Optics*, 6th ed., Pergamon Press, 1980.
4. R. Longhurst, *Geometrical and Physical Optics*, 3rd ed., Longman, 1986.
5. H. Takahasi, S. Suzuki, K. Kato, and I. Nishi, Arrayed waveguide grating for wavelength division multi/demultiplexing with nanometer resolution, *Electron. Lett.* **26**(2): 87–88 (1990).
6. M. Smit and C. van Dam, Phasar-based WDM-devices: Principles design and applications, *IEEE J. Select. Top. Quant. Electron.* **2**(2): 236–250 (1996); also, M. Smit, *Electron. Lett.* **24**(7): 385–386 (1988).
7. K. Takada, H. Yamada, and K. Okamoto, 480 channel 10 GHz spaced multi/demultiplexer, *Electron. Lett.* **35**(22): 1964–1966 (1999).
8. K. Okamoto, K. Moriwaki, and Y. Ohmori, Fabrication of a 64×64 arrayed-waveguide grating multiplexer on Si, *Electron. Lett.* **31**(3): 184–186 (1995).
9. K. Okamoto, K. Syuto, H. Takahashi, and Y. Ohmori, Fabrication of 128-channel arrayed-waveguide grating multiplexer with 25 GHz channel spacing, *Electron. Lett.* **32**(16): 1474–1476 (1996).
10. C. Doerr and C. H. Joyner, Double-chirping of the waveguide grating router, *IEEE Photon. Technol. Lett.* **9**(6): 776–778 (1997).
11. C. Doerr, M. Shirasaki, and C. H. Joyner, Chromatic focal plane displacement in the waveguide grating router, *IEEE Photon. Technol. Lett.* **9**(6): 776–778 (1997).
12. J. Soole et al., Monolithic InP-based grating spectrometer for wavelength-division multiplexed systems at 1.5 μm, *Electron. Lett.* **27**(2): 132–134 (1991).
13. K. McGreer, A flat-field broadband spectrograph design, *IEEE Photon. Technol. Lett.* **7**(4): 397–399 (1995).
14. R. Deri, J. Kallman, and S. Dijaili, Quantitative analysis of integrated optic waveguide spectrometers, *IEEE Photon. Technol. Lett.* **6**(2): 242–244 (1994).
15. Z. Sun, K. McGreer, and J. Broughton, Demultiplexing with 120 channels and 0.29-nm channel spacing, *IEEE Photon. Technol. Lett.* **10**(1): 90–92 (1998).

16. M. Gibbon et al., Optical performance of integrated 1.5 μm grating wavelength multiplexer on InP-based waveguide, *Electron. Lett.* **25**(16): 1441–1442 (1989).

17. *http://www.highwave-tech.com/products/*.

18. I. Nishi, T. Oguchi, and K. Kato, Broad passband multi/demultiplexer for multimode fibres using a diffraction grating with retroreflectors, *IEEE J. Lightwave Technol.* **LT-5**(12): 1695–1700 (1987).

19. J. B. Soole et al., Use of multimode interference couplers to broaden the passband of wavelength dispersive integrated WDM filters, *IEEE Photon. Technol. Lett.* **8**(10): 1340–1342 (1996).

20. *http://www.photonetics.com*.

21. R. Watanabe, K. Nosu, T. Harada, and T. Kita, Optical demultiplexer using concave grating in 0.7–0.9 μm wavelength region, *Electron. Lett.* **16**(3): 106–108 (1980).

22. T. Kita and T. Harada, Use of aberration corrected concave gratings in optical multiplexers, *Appl. Opt.* **22**(6): 819–825 (1983).

23. A. Stavdas, P. Bayvel, and J. E. Midwinter, Design and performance of concave holographic gratings for applications as multiplexers/demultiplexers for wavelength routed optical networks, *Opt. Eng.* (*SPIE*) **35**: 2816–2823 (1996).

24. A. Stavdas et al., The design of a free-space multi/demultiplexers for ultra-wideband WDM networks, *IEEE J. Lightwave Technol.* **19**(11): 1777–1784 (Nov. 2001); also, A. Stavda, *Design of Multiplexer/Demultiplexer for Dense WDM Wavelength Routed Optical Networks*, Ph.D. thesis, Univ. London, 1995.

25. A. Yariv, *Optical Electronics*, HRW International Edition, 3rd ed., 1985.

26. E. Hecht, *Optics*, 2nd ed., Addison-Wesley, Reading, MA, 1987.

27. B. J. Offrein et al., Wavelength tunable optical add-after-drop filter with flat passband for WDM networks, *IEEE Photon. Technol. Lett.* **11**(2): 239–241 (1999).

OPTICAL SIGNAL REGENERATION

GEERT MORTHIER
JAN DE MERLIER
Ghent University — IMEC
Ghent, Belgium

1. INTRODUCTION

It is well known that signals are significantly distorted when they propagate through optical fiber networks (see OPTICAL FIBER COMMUNICATIONS). In modern WDM communications, many channels at different wavelengths and each carrying signals of 10 or even 40 Gbps (gigabits per second) are sent through long stretches of fiber and are traversing several optical amplifiers and optical switches. The propagation of large-bandwidth signals through dispersive optical fiber gives rise to pulse broadening and distortion. The optical amplifiers add noise to the signals and the optical switches can give rise to crosstalk. In large optical networks, the resulting *deterioration* of the signals due to dispersion, noise, and crosstalk may be so large that errorless transmission is no longer possible unless the signals are regenerated at intermediate nodes. This is illustrated in Fig. 1, which schematically shows the increase in bit error rate (BER) as a function of the distance in an optical transmission link with and without *regeneration*.

Regeneration can be achieved by converting the optical signals into electronic signals (using optical receivers), regenerating the electronic signals, and by retransmitting the signals optically (i.e., using a laser). However, because of the involvement of power electronics, this *optoelectronic regeneration* becomes more and more costly as the bit rate increases. All-optical regeneration becomes a very interesting alternative to optoelectronic regeneration in this case. The fact that *all-optical regeneration* can in principle be extended to multichannel regeneration is, for WDM systems, especially attractive.

Different types of optical regenerators, based on fiber loops or on *photonic integrated circuits* (PICs), have been proposed so far in the literature. The regenerators based on PICs are in general much more stable than the ones based on fiber loops (which require, e.g., *polarization control*). The latter can be operated at higher speeds though. Fiber loop regenerators and high-speed regenerators in general are usually aiming at *return-to-zero* (*RZ*) *signals*. Regeneration of *non-return-to-zero* (*NRZ*) *signals* is often limited to lower bit rates.

2. FIGURES OF MERIT

One typically classifies regeneration into 1R, 2R, and 3R regeneration; 1R consists of simple reamplification, 2R includes *reamplification* and *reshaping*, and 3R includes reamplification, reshaping, and *retiming*. Figure 2 illustrates 3R regeneration. It is generally believed that 2R regeneration may suffice at bit rates below 10 Gbps. From 10 Gbps on, optical regeneration could consist of 3R regeneration at selected nodes and 2R regeneration at other nodes. At these very high bit rates, *timing jitter* or, in other words, fluctuations in the duration of the individual bits (see OPTICAL NOISE) can be a serious cause of degradation and must be alleviated by retiming techniques.

A number of properties determine the performance of regenerators. The quality of the *regeneration* itself is characterized by the extinction ratio improvement

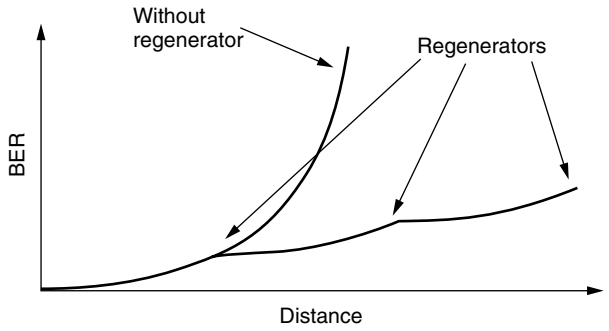


Figure 1. Increase of BER versus link distance with and without regeneration.

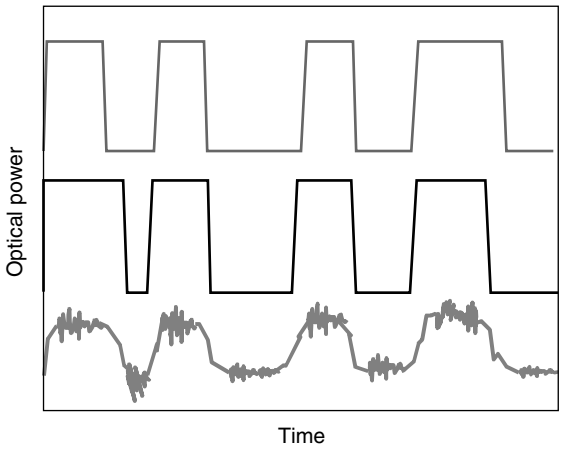


Figure 2. Schematic illustration of 3R optical regeneration; lower trace — signal before regeneration; middle trace — signal after 2R regeneration; upper trace — signal after 3R regeneration.

(see OPTICAL MODULATION) and by the *noise reduction* (determined by the flatness of the output vs. input power characteristic for both spaces and marks). An ideal 2R regeneration or *decision characteristic* as well as a more realistic characteristic are shown in Fig. 3. The *extinction ratio improvement* can be defined as $(P_{2,\text{out}}/P_{1,\text{out}})/(P_{2,\text{in}}/P_{1,\text{in}})$. In addition, the maximum bit rate should be as high as possible, while the required input power levels to the regenerator should be low or modest. The minimum required input power (for a digital one) is related to the decision threshold.

Finally, the ability to use the same device for regeneration over a broad wavelength range, even if one or more current settings have to be changed, is considered as a major advantage for WDM applications.

3. 2R REGENERATION

In order to obtain optical 2R regeneration, the nonlinear effects in optical waveguides (see OPTICAL WAVEGUIDES) are exploited. The incident light is used to influence the refractive index of the material through which it propagates. This happens via the interaction with the charge carriers in semiconductor waveguides. The refractive index variation in turn affects the propagation delay in the waveguides or, in other words, the phase of

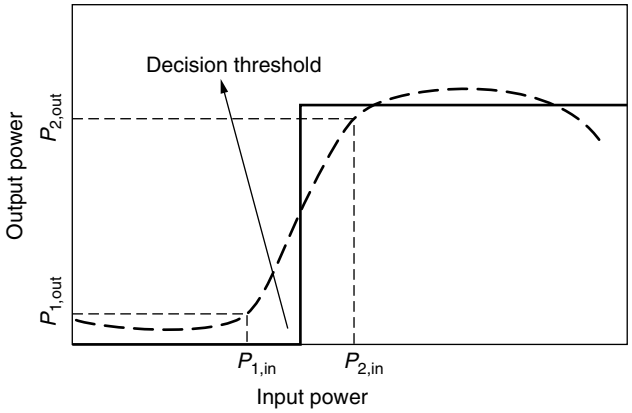


Figure 3. Ideal (full line) and realistic (dashed line) decision characteristic.

the optical field at the other end of the waveguide. The resulting phase changes can cause drastic changes in the output of a component when the waveguides are used in an *interferometric layout*.

3.1. SOA-Based Regenerators

The most promising results, in terms of regeneration, have been achieved with interferometric structures, such as the *Mach–Zehnder interferometer* (MZI) or Michelson interferometer (MI), containing a *semiconductor optical amplifier* (SOA) (see OPTICAL AMPLIFIERS) in each arm. With these interferometric structures, shown in Fig. 4, regeneration can be obtained through simultaneous *wavelength conversion* (see OPTICAL FREQUENCY CONVERSION) or without wavelength conversion, using a passthrough scheme.

When regeneration is performed using a passthrough scheme, the currents in both SOAs are chosen such that destructive interference is obtained at low input powers. Because of the difference in current, the saturation behavior of both SOAs will also differ. As a result, the phase difference between both arms will be modulated by the power variation in the data signal. This phase modulation is converted to an amplitude modulation at the output of the interferometer. Regeneration (although mainly for logical '1s') without wavelength conversion has been demonstrated at 40 Gbps using the Mach–Zehnder interferometer. When using a *Michelson interferometer*, input and regenerated signal have to be separated using a circulator.

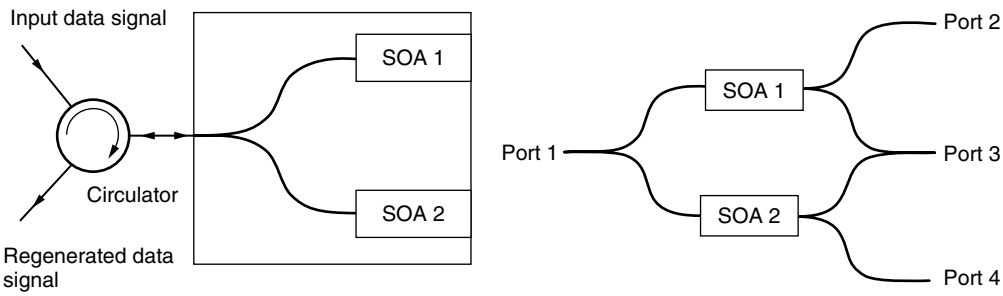


Figure 4. Michelson interferometer–based regenerator (left) and Mach–Zehnder interferometer–based regenerator (right).

The highest speed can theoretically be achieved with Mach–Zehnder interferometers and with simultaneous wavelength conversion. Wavelength conversion in the codirectional scheme (Fig. 4) is obtained by injecting a CW signal in port 3 and injecting the data signal in port 2. The electron density in SOA 1 is then modulated by the data signal. This causes a change in the phase difference between the arms, which causes modulation of the power in the CW signal at the output of the MZI. The CW signal can also be coupled into port 1, in a *counterdirectional scheme*. This allows the in- and output wavelengths to be the same and avoids the need for a filter at the output. Because of the sinusoidal dependence of the output on the phase difference between the arms, the polarity of the converted signal can be maintained or converted as compared to the incident data signal.

This scheme can be used at bit rates up to 40 Gbps for RZ signals. For the very high bit rates, however, a RZ input signal is converted to a NRZ signal [1], due to the finite lifetime of the electron density. Therefore other schemes are used at these very high bit rates.

Differential delay techniques make specific use of the fact that the interferometer output depends on the phase difference. In this scheme (Fig. 5), the data signals are injected in the SOAs in both arms of the MZI with a small delay between both pulses. The induced phase change in the arm where the data signal is injected first (e.g., ϕ_1), can be canceled out once the other pulse is injected into the other arm (giving a phase change ϕ_2). This differential control scheme has been shown feasible at bit rates of 40 Gbps. A filter is obviously required at the output to suppress the signal at λ_1.

Other SOA-based regenerators have been proposed on the basis of the nonlinear properties of, e.g., an active *multimode interference* (MMI) *coupler* and an active directional coupler. Both devices have been verified only by static measurements of the transfer characteristic, but for this static regime (and hence also for the low-bit-rate regime) they exhibited improved regeneration characteristics as compared to the MZI- and MI- based regenerators [e.g., 2].

3.2. Nonlinear Optical Loop Mirrors

Optical regenerators based on the *nonlinear optical loop mirror* (NOLM) make use of the ultrafast but weak Kerr nonlinearity (Fig. 6; see NONLINEAR EFFECTS IN FIBERS). This layout has the disadvantage that very high powers and several hundreds of meters of dispersion-shifted fiber (DSF) are required. The data signal is coupled in and out of the NOLM using a wavelength-dependent coupler (WDC). A CW-signal enters the NOLM through port 1 of a 50–50 coupler. Therefore, 50% of the CW signal propagates clockwise and 50% of it propagates counterclockwise in the ring. If a data pulse is present, the CW signal copropagating with the pulse experiences a nonlinear phase shift that is proportional to the data pulse intensity. At the output of port 2, one obtains interference between the phase-modulated clockwise and the counterclockwise propagating CW light. The light at the CW wavelength that is coupled out of the ring at port 2 is therefore a regenerated version of the data signal. The regenerative capabilities of the NOLM have been demonstrated at 80 Gbps, but this NOLM-based regenerator should be capable of regeneration at bit rates well over 100 Gbps [3].

3.3. Electroabsorption Modulators

Nonlinear behavior in a reverse-biased *electroabsorption modulator* (EAM) is achieved by using an intense input optical pulse to produce a large number of photogenerated charged carriers in the highly absorptive waveguide. Drift and diffusion of these carriers distort the electric field and cause a reduction of the field. As a result of the reduced electric field, the absorption decreases and a

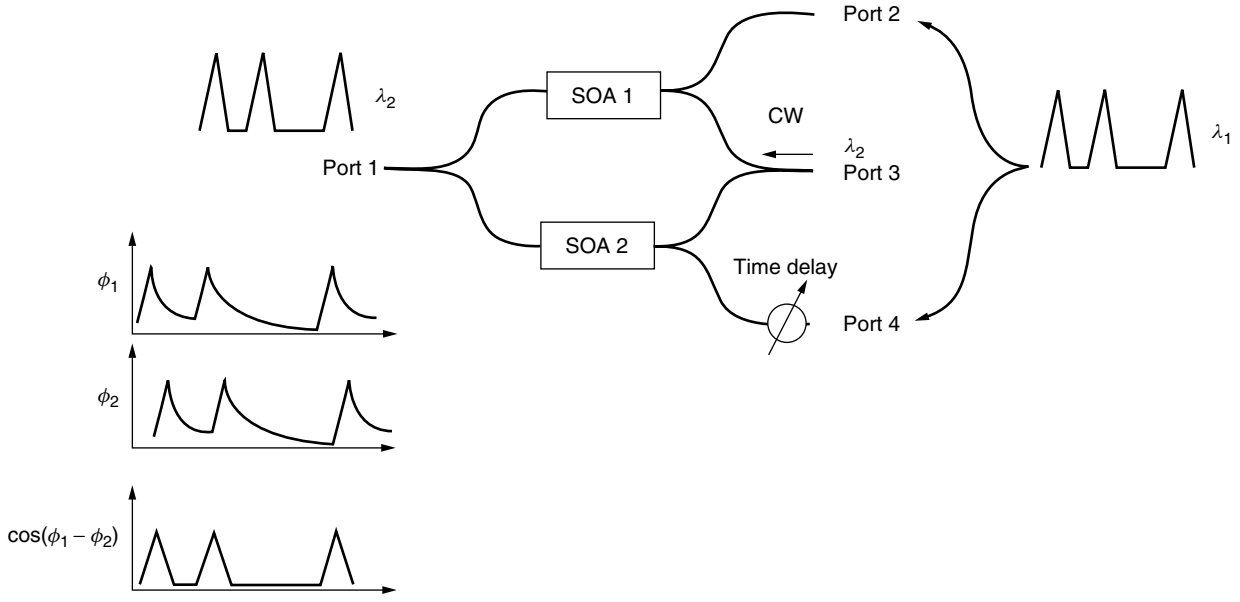


Figure 5. Differential delay scheme for a MZI-based regenerator.

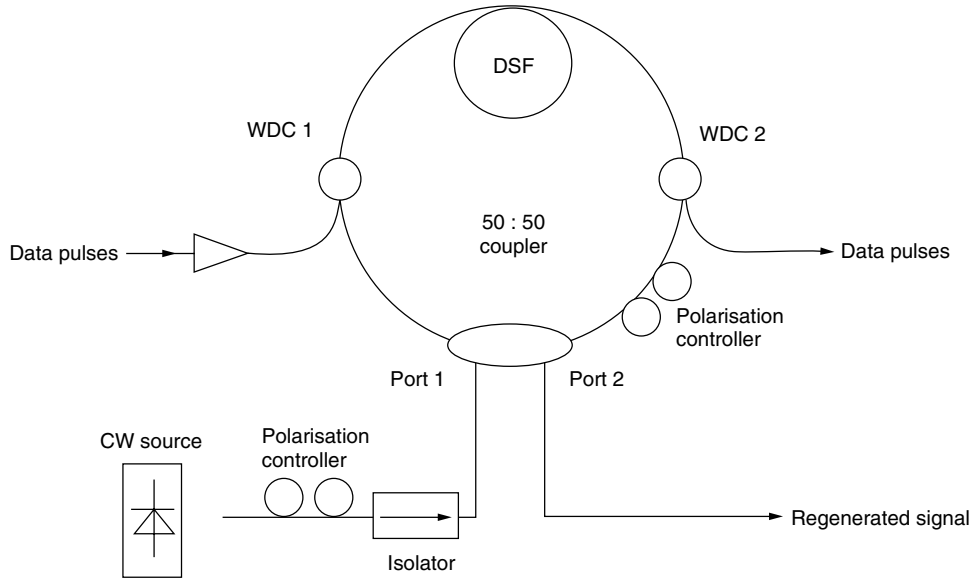


Figure 6. Scheme of nonlinear optical loop mirror (NOLM)-based regenerator.

transmission window is created for the pulse. When an intense optical pulse propagates through the EAM, a CW probe signal traversing the waveguide experiences the transient increase in transmission followed by a fast resumption of the absorption. As a result, a short pulse is generated at the wavelength of the probe signal. Weak input signals are absorbed without changing the transmission state of the EAM. This type has been demonstrated at bit rates of 40 Gbps.

3.4. Q-Switched Lasers

The *Q-switched laser* consists of three sections: two Bragg sections and one integrated phase tuning section Fig. 7. The second Bragg section is biased near transparency and is used as a dispersive reflector. The different sections are designed and biased such that lasing is achieved in the laser section only when a matched feedback is obtained from the reflector section.

Injection of a high-power data pulse changes the refractive index in the reflector section, which, in turn, causes a shift in the reflection band away from the lasing wavelength (change of Q factor), stops the required feedback and eventually ends the lasing in the laser section. Lasing starts again when the power in the signal falls again. Optimization of this scheme has led to successful tests up to 10 Gbps.

4. OPTICAL CLOCK EXTRACTION AND 3R REGENERATION

3R regenerators consist of 2R regenerators in combination with *optical clock extraction*. The periodic optical clock is then typically modulated in a more or less digital manner by the 2R-regenerated signal. Optical clock extraction itself is generally based on the use of self-pulsating laser diodes of which the self-pulsation frequency locks to the bitrate of the incoming signal. The self-pulsating laser diode can be either a *mode-locked laser* [4] or a multisection DFB laser [5]. Clock recovery up to 40 Gbps has been demonstrated with both types of devices.

The scheme for clock extraction based on a multisection DFB laser is the same as the scheme of Fig. 7, except that the regenerated signal is replaced by a clock signal. The DFB laser (see OPTICAL TRANSMITTERS AND RECEIVERS) now has to be designed such that lasing occurs in a self-pulsating regime and at a wavelength far away from the wavelength of the injected signal. The self-pulsation frequency of the free-running laser is determined by the injected currents

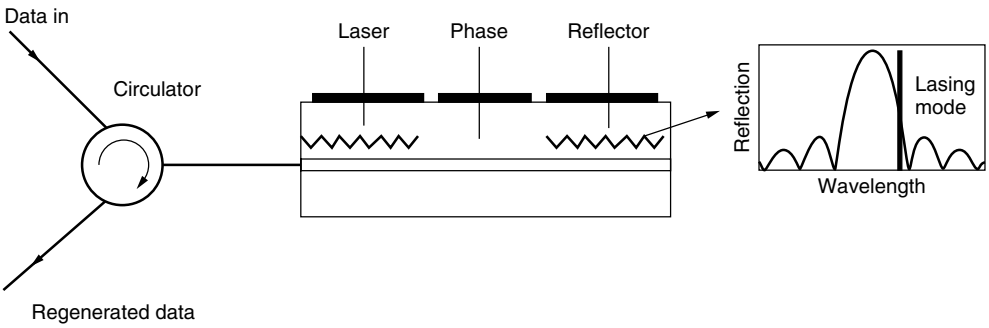


Figure 7. 2R regenerator based on a Q-switched laser.

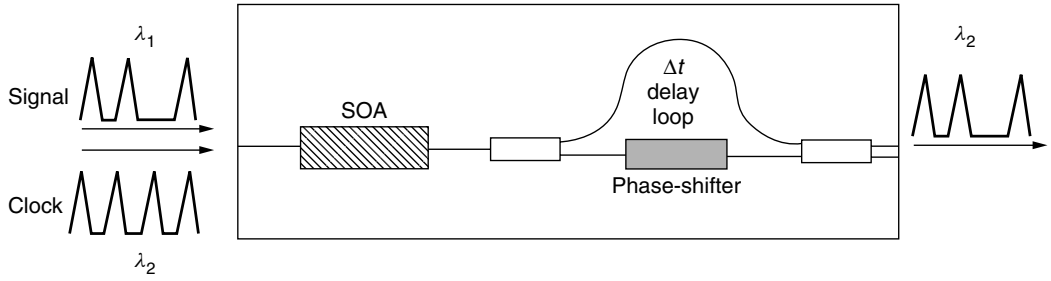


Figure 8. Scheme of semiconductor optical amplifier delayed interference device.

and can typically vary from 5 to over 40 GHz. When a data signal with a bit rate close to the repetition rate of the self-pulsation is injected, the self-pulsation synchronizes to the data signal and assumes a frequency (in GHz) that is exactly the bit rate (in Gbps).

All-optical 3R regeneration has been proposed by various groups [4–6]. In all cases, only a RZ signal format is considered. A recent idea is the use of a SOA *delayed interference device*, shown in Fig. 8. The delay Δt matches the period of the clock. The delayed interference coupler therefore produces at its output the interference of each clock pulse with the previous clock pulse. Pulses are produced in the lower arm of the coupler in the case of constructive interference and in the upper arm in the case of destructive interference. However, the phase and amplitude of the clock pulses are modulated in the SOA by the power of the data signal pulses. If the power of the data pulses is chosen carefully, a phase shift of π in the clock pulses can be obtained after each data pulse. Hence, each data pulse results in a constructive interference at the coupler and thus a clock pulse in the lower output of the coupler.

Optical 3R regeneration for NRZ signals has not been reported so far.

5. MULTICHANNEL REGENERATION

Multichannel optical regeneration is generally devised as a combination of phased-array multiplexers and demultiplexers and an array of single-channel regenerators. A possible concept for multichannel 2R regeneration, for instance, is depicted schematically in Fig. 9. The channels are demultiplexed by the *phased array* (see Optoelectronic Devices), and individual channels are fed to active Michelson interferometers [3]. After regeneration (and reflection), the channels are multiplexed again in the phased array. At the output waveguide of the phased array a circulator guarantees that the regenerated WDM signal is routed in another direction than that of the incoming signal.

Multichannel 3R regeneration requires simultaneous all-optical clock recovery of all WDM channels. This has been demonstrated recently using a module consisting of a demultiplexer, an array of SOAs (see Optical Amplifiers) and a multiplexer, all placed in an actively mode-locked fiber ring laser configuration [7]. The clock recovery module is shown in Fig. 10. The circuit of Fig. 10 forms an actively mode locked laser for each incoming channel. The mode locking results from the injection of a datastream into each SOA, which causes amplitude and

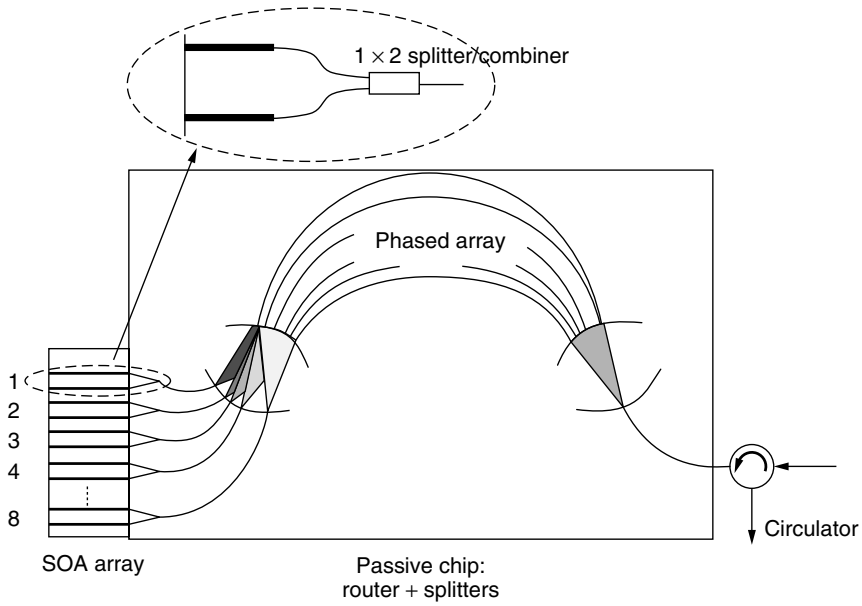


Figure 9. Top view of a multichannel 2R regenerator.

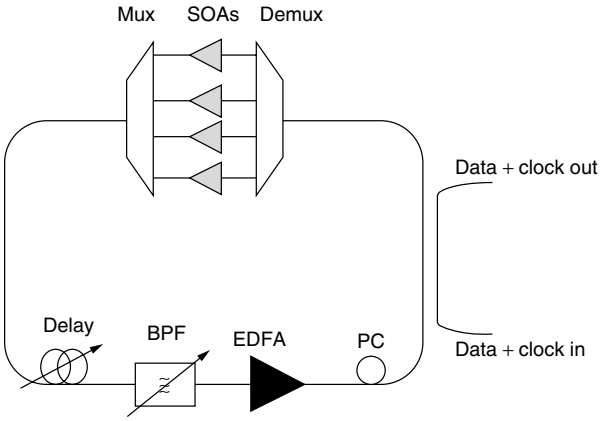


Figure 10. Multichannel clock recovery circuit [6] (PC — polarization control, EDFA — erbium-doped fiber amplifier, BPF — band-pass filter, Mux and Demux — phased arrays).

phase modulation of the light. The presence of the *tunable band-pass filter* allows the use of different phased array orders for the data and the extracted clock.

BIOGRAPHIES

Geert Morthier received the M.S. and Ph.D. degrees in electrical engineering from the University of Ghent, Belgium, in 1987 and 1991, respectively. He joined IMEC in 1991 and has been a group leader since 1992. Since 2001, he is also parttime professor at the University of Ghent. He has been author or co-author of approximately 100 publications and holds 5 patents. His areas of interest are DFB and tunable laser diodes and optical signal processing.

Jan De Merlier received the degree in physical engineering in 1998 at Ghent University, Belgium, and is currently working toward a Ph.D degree in electrical engineering at the Department of Information Technology, Ghent University. His main research interests are the dynamic properties of semiconductor optical amplifiers for use in all-optical regenerators.

BIBLIOGRAPHY

1. J. Leuthold et al., 100 Gbit/s all-optical wavelength conversion with integrated SOA delayed-interference configuration, *Electron. Lett.* **36**: 1129–1130 (2000).
2. J. De Merlier et al., Experimental demonstration of 15 dB extinction ratio improvement in a new 2R optical regenerator based on an MMI-SOA, *27th Eur. Conf. Optical Communication* (ECOC'2001), Amsterdam, Sept. 2001.
3. *J. K. Lucek, K. Smith, "All-optical signal regenerator" Optics Letters, vol 18, Aug. 1993, pp. 1226–1228.*
4. C. Bornholdt et al., Self-pulsating DFB laser for all-optical clock recovery at 40 Gbit/s, *Electron. Lett.* **36**: 327–328 (2000).
5. D. T. K. Tong et al., 160 Gbit/s clock recovery using electroabsorption modulator-based phase-locked loop, *Electron. Lett.* **36**: 1951–1952 (2000).
6. J. Leuthold et al., Novel 3R regenerator based on semiconductor optical amplifier delayed interference configuration, *IEEE Photon. Technol. Lett.* **13**: (2001).
7. V. Mikhailov and P. Bayvel, Multiwavelength all-optical clock recovery using an integrated semiconductor optical amplifier array, *Proc. ECOC'2000*, Munich, Sept. 2000, Vol. 3, pp. 63–64.

OPTICAL SOLITONS

Magnus Karlsson
Peter Andrekson
Chalmers University of Technology
Göthenburg, Sweden

1. INTRODUCTION

A traveling wave that is localized in the sense that it does not spread its energy while propagating, is defined as a *solitary wave*. A *soliton* is a solitary wave with the additional property that it can collide with other solitons and emerge unaffected with respect to energy, shape and momentum after the collision. Over the years, however, it has become common practice, especially in optics, to use "soliton" also in the less strict definition, although the mathematical differences between solitons and solitary waves are profound.

In 1834, the Scottish engineer John Scott Russel made the first scientific observation of a soliton, in the form of a single water wave that rolled forward in a canal with unchanged velocity and shape. The phenomenon was mathematically explained sixty years later in terms of a solution to a nonlinear partial differential equation that governs the motion of shallow water waves. Further progress on soliton physics came in the 1960s, when novel analytic methods were developed to exactly solve equations governing solitons.

It was realized in 1973, by Hasegawa and Tappert [1] that the weak Kerr nonlinearity present in optical fibers (which makes the refractive index increase in proportion to the optical intensity) might counteract the pulse broadening induced by group velocity dispersion (GVD). In fact, the two effects can form a stable balance in the form of a soliton pulse, which then propagates without changing shape.

Because of the lack of short-pulse laser sources at wavelengths above 1.3 μm, and low-loss silica fibers, it took another seven years for optical soliton pulses to be experimentally verified. In an experiment by Mollenauer et al. [2] in 1980, soliton pulse transmission over 700-m fiber was demonstrated. During the 1980s the soliton research aimed toward the use of solitons as information carriers in optical communications, and in 1990 the first data transmission experiment using solitons [2.8 Gbps (gigabits per second) over 23 km of fiber] were reported by Iwatsuki et al. [3].

During the 1990s, soliton-based communication systems have matured, an important reason being the development of the erbium-doped fiber amplifier (EDFA), which made high power levels commercially available. The most recent development has been toward the use of solitons in alternating dispersion maps (so called *dispersion management*), and such systems have reached performance levels near commercialization.

In the present review, we discuss both theoretical and experimental aspects of soliton transmission. We will distinguish between *conventional solitons*, which have constant dispersion during the transmission, and *dispersion managed solitons* (although solitary waves would be the proper mathematical name) for which the dispersion varies periodically during transmission.

2. WAVE PROPAGATION IN OPTICAL FIBERS

We restrict the treatment here to forward-going (in the z direction) waves along single-mode fibers. This means that the lightwave propagates according to $\exp(-i\beta z)$, where the wavenumber $\beta = n\omega/c$ is a function of the angular frequency ω, the refractive index n, and the vacuum speed of light c. In fused silica fibers the refractive index depends on the frequency and intensity of the light, and it is usually modeled as

$$n = n_0(\omega) + n_2|E|^2 \tag{1}$$

where $n(\omega)$ is the frequency-dependent part, n_2 is the nonlinear part, and E is the electric field of the wave. This nonlinear dependence on the wave intensity is known as the *optical Kerr effect*. In fused silica, the Kerr coefficient n_2 is very small; on the order of 10^{-20} m^2/V^2. The nonlinear part of the refractive index will enter as a nonlinear part in the dispersion relation, which can be written as $\beta = \beta_{\text{lin}}(\omega) + \beta_{\text{nonlin}}$. After averaging the nonlinear index over the mode profile, the nonlinear part becomes

$$\beta_{\text{nonlin}} = \frac{2\omega Z_0 n_2}{c n_0 A_{\text{eff}}}|u|^2 = \gamma|u|^2 \tag{2}$$

where $Z_0 \approx 120\pi\,\Omega$ is the wave impedance of vacuum, n_0 is the average refractive index, and A_{eff} is the effective mode area of the fiber. All these constants are contained in the nonlinear fiber coefficient γ, which has units of m^{-1} W^{-1}. We also use the wave amplitude u of the light, normalized so that the transmitted power is given by $|u|^2$. The value for γ varies from different fiber types, due to the dependence on the core area A_{eff}, but for standard single-mode fibers at $\lambda = 15.50$ nm it is approximately 2.2 m^{-1} W^{-1}.

2.1. Linear Dispersive Fiber Transmission

Linear transmission is straightforward to carry out in the frequency domain, where the wave amplitude spectrum $\tilde{u}(z,\omega)$ propagates according to $\tilde{u}(z,\omega) = \exp(-iz\beta_{\text{lin}}(\omega))\tilde{u}(0,\omega)$. This can also be expressed in the time domain, but then the function $\beta_{\text{lin}}(\omega)$ should be interpreted as the operator $\beta_{\text{lin}}(-i\partial/\partial t)$. In general we can Taylor expand the dispersion relation $\beta_{\text{lin}}(\omega)$ around the carrier frequency ω_0 to arbitrary orders. The first two terms of this expansion will correspond to the phase and group velocities of the wave, and can be removed with a suitable choice of coordinate system. Thus we can write

$$\beta_{\text{lin}}(\omega) = \frac{\beta_0''}{2}\omega^2 + \frac{\beta_0'''}{6}\omega^3 + \cdots \tag{3}$$

where now ω is measured from the carrier wavelength ω_0. The coefficient β_0'' is known as the *group velocity dispersion* (GVD) *coefficient*. The GVD is said to be *normal* if v_g decreases with frequency (i.e., $\beta_0'' > 0$), and *anomalous* if v_g increases with frequency (i.e., $\beta_0'' < 0$). Another common measure of the fiber dispersion is the *dispersion parameter*, D. The dispersion parameter D [ps/(km · nm)] is related to the GVD coefficient β_2 [ps^2/km] as $D = -\beta_2 2\pi c/\lambda^2$. In standard single-mode fibers (SMFs) the GVD is to a good approximation a linear function of the frequency, with slope $\beta_0''' \approx 0.1$ ps^3/km and zero at the *zero-dispersion wavelength* λ_0, at which $\beta_0'' = 0$. Above (below) this wavelength we have anomalous (normal) GVD. In standard single-mode fibers $\lambda_0 \approx 1.33\,\mu m$, whereas in dispersion-shifted fibers (DSFs) $\lambda_0 \approx 1.55\,\mu m$. The fiber manufacturing of today has reached a very mature level, and it is possible to tailor fibers to have a wide range of dispersion zeros and dispersion slopes.

It is instructive to consider the linear evolution of the Gaussian pulse $u(0,t) = u_0\exp(-t^2/2T_0^2)$, which is

$$u(z,t) = u_0\frac{T_0}{\sqrt{T_0^2 + i\beta_0'' z}}\exp\left(-\frac{t^2}{2(T_0^2 + i\beta_0'' z)}\right) \tag{4}$$

and where β_0''' has been neglected. The pulsewidth is determined by the real part of the exponent, which means that the width will broaden according to $T(z) = \sqrt{T_0^2 + (\beta_0'' z/T_0)^2}$. The imaginary part of the exponent corresponds to a phase modulation over the pulse, of the form $\exp(-iC(z)t^2)$, which does not affect the pulse width. Instead, this phase modulation (known as the *linear chirp* of the pulse) shows that the frequency components are redistributed within the pulse so that the slow and fast frequency components are put in the respective trailing or leading edges of the pulse. This is evident by considering the instantaneous frequency $\omega_i(t)$ of the pulse which is defined as $\omega_i(t) = d(\arg(u))/dt$, and for the Gaussian pulse it will be $\omega_i(t) = -2tC$, which reveals that the trailing (leading) part of the pulse is red (blue)-shifted for positive C (i.e., anomalous dispersion). For normal dispersion it is the other way around.

2.2. Nonlinear Fiber Transmission: Self-Phase Modulation

If we neglect the linear parts in the dispersion relation for the moment and concentrate on the nonlinear part, we see that an initial pulse $u(0,t)$ will propagate according to $u(z,t) = u(0,t)\exp(-i\gamma z|u(0,t)|^2)$. This is called *self-phase modulation* (SPM) as the pulse modulates its own phase. The corresponding instantaneous frequency will be $\omega_i(t) = -\gamma z d(|u|^2))/dt$, which is negative in the leading edge and positive in the trailing edge of the pulse, which is the same as for normal dispersion. As a result the nonlinearity will increase the dispersive spreading for normal dispersion, and decrease it for anomalous dispersion. In the case when the nonlinearity and the dispersion exactly cancel out, an optical soliton that propagates without dispersive spreading is formed.

3. CONVENTIONAL FIBER SOLITONS

3.1. The Nonlinear Schrödinger Equation

Taking both self-phase modulation and dispersion into account, we get the following propagation equation valid

for light in an optical fiber:

$$i\frac{\partial u}{\partial z} + \frac{\beta_0''}{2}\frac{\partial^2 u}{\partial t^2} - \gamma|u|^2 u = i\frac{\alpha}{2}u + i\frac{\beta_0'''}{6}\frac{\partial^3 u}{\partial t^3} \tag{5}$$

where $1/\alpha$ is the fiber loss length, which is of the order of 25 km at $\lambda = 15.50$ nm. For fiber lengths less than 1 km this can be ignored. Retaining only the terms on the left-hand side, gives the *nonlinear Schrödinger* (NLS) *equation*, which was derived for fibers originally by Hasegawa and Tappert in 1973 [1]. The NLS equation is a universal nonlinear propagation equation in the sense that it arises in many different fields of physics, and it has, for nonlinear partial differential equations the unusual and nice feature that it is *integrable*, which means that its initial-value problem can be solved exactly.

It is the anomalous dispersion case that is of most interest to us, and we will summarize the properties of the NLS equation in this case. For this it is convenient to work with the NLS equation in normalized form, and we introduce the "soliton normalizations"

$$q = u\sqrt{\frac{\gamma}{L_d}} \quad \tau = \frac{t}{t_0} \quad \xi = \frac{z}{L_d} \equiv \frac{z|\beta_0''|}{t_0^2} \tag{6}$$

where t_0 is the pulse width and L_d the dispersive length. In the case of anomalous dispersion, the NLS equation in these normalized units becomes

$$i\frac{\partial q}{\partial \xi} = \frac{1}{2}\frac{\partial^2 q}{\partial \tau^2} + |q|^2 q \tag{7}$$

The mathematical theory for the solution of this equation, which demonstrates the *inverse scattering transform* (IST) for the NLS, was given 1972 in an important paper by Zakharov and Shabat [4]. They found that a crucial role is played by the *soliton* solution to Eq. (7)):

$$q_{\mathrm{sol}}(\xi,\tau) = A\,\mathrm{sech}(A(\tau - V\xi)) \times \exp\left[-iV\tau + i\frac{(V^2 - A^2)\xi}{2}\right] \tag{8}$$

where A is an arbitrary soliton amplitude and V is an arbitrary frequency shift. Note that the soliton moves with the "velocity" V in the retarded reference frame due to this shift, which means that the soliton moves with the group velocity of the carrier wavelength. Because the soliton does not broaden dispersively, it seems very attractive for information transfer.

The IST reveals that all solutions to the NLS equation consist of solitons and dispersive radiation [i.e., dispersively broadening pulses like in Eq. (4)]. Early numerical simulations [1] also demonstrated the stable-attractor properties of the soliton. The theory also shows that any number (say, N) solitons can be present simultaneously, forming an "N soliton" solution. This can be either N well-separated soliton pulses, or if the pulses are clumped together, an oscillating N-soliton structure, called a *breather*.

One important implication of the IST is that we can already from the initial condition $q(0,\tau)$ conclude what kind of soliton will emerge for large values of ξ. For the initial condition $q(0,\tau) = A\mathrm{sech}(\tau)$, the emerging soliton is an N : th order soliton, where $A = N + \eta$, N is an integer and $|\eta| < \frac{1}{2}$. In the case where the initial pulse is given by $(1+\eta)\mathrm{sech}(\tau)$, an emerging soliton with $A = 1 + 2\eta$ is formed [5]. For an arbitrarily shaped real initial pulse, the condition for soliton creation is that the pulse area exceeds $\pi/2$ in normalized units. Note that the soliton area, $\int_{-\infty}^{+\infty} q_{\mathrm{sol}} d\tau = \pi/2$ is independent of the free parameters A and V, so that there is a critical *area* for soliton creation, rather than a critical power level. Therefore, solitons having all power levels $|q(0,0)|^2 = A^2$, and energies $\int |q_{\mathrm{sol}}|^2 d\tau = 2A$ exist. For a given pulse duration and shape, however, the condition for creation gives the necessary power to obtain a soliton, and this can be viewed as a critical power level.

Finally it is instructive to transform back to physical parameters and write the fundamental soliton of duration t_0 as

$$u_{\mathrm{sol}}(z,t) = \sqrt{\frac{|\beta_0''|}{\gamma t_0^2}}\mathrm{sech}\left(\frac{t}{t_0}\right)\exp[-iz\beta_{\mathrm{sol}}] \tag{9}$$

where the soliton wavenumber $\beta_{\mathrm{sol}} = |\beta_0''|/2t_0^2 = (2L_d)^{-1}$. It should be emphasized that the soliton wavenumber β_{sol} must lie in a regime where it cannot equal linear wavenumbers. For the NLS equation we have $\beta_{\mathrm{lin}} = -|\beta_0''|\omega^2/2 < 0$ and $\beta_{\mathrm{sol}} > 0$. Examples of cases when this condition is not fulfilled are when higher-order dispersion or periodic amplification is present. The soliton peak power P_s [W] and energy E_s [J] can be expressed as $P_s = |u_{\mathrm{sol}}(z,0)|^2 = |\beta_0''|/\gamma t_0^2$ and $E_s = 2P_s t_0 = 2|\beta_0''|/\gamma t_0$.

Next, we will review the effects of a few of the most important perturbations of solitons in fibers, and see how the right-hand side terms of Eq. (5) affect soliton propagation.

3.2. Solitons in Presence of Third-Order Dispersion

The significance of third-order dispersion (3OD) depends on the amount of soliton energy that lies in the normal-dispersion regime, which in practice depends on the pulsewidth and the carrier wavelength. The effects from 3OD are particularly important near the zero-dispersion wavelength, and the governing equation is then

$$i\frac{\partial q(\xi,\tau)}{\partial \xi} = \frac{1}{2}\frac{\partial^2 q}{\partial \tau^2} + |q|^2 q + i\varepsilon\frac{\partial^3 q}{\partial \tau^3} \tag{10}$$

where $\varepsilon = \beta_0'''/6|\beta_0''|t_0$. Using perturbation theory, it is possible to find the lowest-order (in ε) corrections to the soliton $q = A\mathrm{sech}(A\tau)\exp[-i\xi A^2/2]$, which reveals that the soliton will be spectrally shifted into the anomalous dispersion regime an amount εA^2, and as a result acquire a new group velocity. However this is not the whole story.

We note that the linear dispersion relation when third-order dispersion is present is $\beta_{\mathrm{lin}} = -|\beta_0''|\omega^2/2 + \beta_0'''\omega^3/6$, and the soliton wavenumber is $\beta_{\mathrm{sol}} = |\beta_0''|/2t_0^2$. Evidently, at a certain frequency β_{sol} equals β_{lin}, and for that frequency the soliton will act as a source for linear waves, which will then radiate and leave the soliton. The radiating frequency ω_r is approximately equal to $(2t_0\varepsilon)^{-1}$ when ε is small. This frequency lies in the normal dispersion regime. Since it is the soliton that acts as a

source for the radiation, the amplitude of the radiation will be proportional to the soliton spectral amplitude at ω_r, which is $u_{\rm sol}(t_0\omega_r) \sim \mathrm{sech}(\pi t_0\omega_r/2) \sim \exp(-\pi t_0\omega_u/2) = \exp(-\pi/(4\varepsilon))$ [e.g., 6]. Thus the radiation is exponentially small for small ε, but the fact that it cannot be expressed as a Taylor series in ε makes conventional perturbation analysis very difficult [7]. In communications, this kind of radiation must be avoided, and the solution is to make sure the soliton is sufficiently located to the anomalous dispersion regime, specifically, that ε is small enough. As a design criterion $\varepsilon < 0.04$ has been suggested [7].

3.3. Solitons in Presence of Amplification and Loss

The most important property that has been neglected in the derivation of the NLS equation for optical pulses is the effect of loss. We base the discussion on Eq. (5), retaining only the loss term $\alpha/2u$ on the right-hand side. For a soliton solution $q_{\rm sol} = A\mathrm{sech}(A\tau)\exp[-i\xi A^2/2]$, one can show that the amplitude and width are adiabatically modified according to $A(\xi) = A_0\exp[-\alpha L_d\xi]$. However, after an initial phase this rate of pulse broadening and amplitude decay will be replaced by conventional linear dispersive broadening and amplitude decay [8].

In soliton communication systems the effect of fiber loss must be compensated for by periodically spaced amplifiers. This will introduce a periodic perturbation on the soliton, with period equal to the amplifier distance L_a, and with a corresponding perturbation wavenumber $\beta_{\rm per} = 2\pi/L_a$. This wavenumber can make up for the wavenumber difference between the linear waves and solitons, and if the equation $\beta_{\rm sol} = \beta_{\rm per} + \beta_{\rm lin}(\omega)$ has any solutions, the corresponding frequencies will be unstable and radiate. From the condition above we find the radiating frequencies ω_r as $1 + (t_0\omega_r)^2 = 4\pi L_D/L_a$, and since the soliton spectral width is of the order t_0^{-1}, it suffices that $t_0\omega_r \gg 1$, and usually $L_D > L_a$ is adopted as a design criterion. This fact, that the dispersive length must be larger than the amplifier spacing is a serious obstacle for conventional soliton systems at high bit rates, and as a result solitons in the early 1990s showed most success in transoceanic systems, where the total system length is very long, but where the data rate is relatively moderate.

Finally it should be emphasized that the launched peak power of the soliton at each amplifier should be such that the path-average power between the amplifiers equals the soliton power, P_s. Since the peak power $P_{\rm peak}$ falls off as $\exp(-\alpha z)$, the average peak power over one amplifier span is $P_{\rm peak}(1-\exp(-\alpha L_a))/\alpha L_a = P_{\rm peak}(G-1)/G\ln(G)$, where $G = \exp(\alpha L_a)$ is the amplifier gain.

3.4. Sources of Timing Jitter

Another transmission obstacle is the various sources of random movement in the bit slot, namely, timing jitter of the pulses in the data transmission link. There are various sources of timing jitter, such as soliton interaction, Gordon–Haus, acoustic, and WDM–collision-induced types of jitter.

3.4.1. Soliton Interactions. Because solitons are nonlinear pulses, they will *interact* with adjacent pulses in the pulsetrain, as pointed out quite early [9]. Interaction between solitons of the same polarization and wavelength is *phase-sensitive*, so in-phase solitons will attract each other whereas out-of-phase solitons will repel each other. The interaction can be quantified via the collapse distance z_c at which in-phase solitons merge, and one can show that $z_c \approx \exp(T/2t_0)L_D\pi/4$, where T is the bit separation. The relative pulse separation is usually defined as T/T_{FWHM}, where $T_{FWHM} = 1.76t_0$ is the soliton width [in the full-width half-maximum (FWHM), sense]. A typical separation used in experiments is $T/T_{FWHM} \approx 5$.

Orthogonally polarized solitons interact substantially less, since it is then the intensity overlap that causes the interaction, rather than the amplitude overlap as for copolarized solitons. Polarization-multiplexed solitons, where adjacent pulses have orthogonal polarization can therefore be packed almost twice as dense, and $T/T_{FWHM} \approx 2.5$ is a commonly used separation.

3.4.2. Gordon–Haus Jitter. The noise from the inline amplifiers will give rise to a small jitter in the carrier frequency of each soliton, thereby changing the group velocity and hence affect the arrival time of each pulse. This is known as the *Gordon–Haus effect* after the authors of Ref. [10], and it has to be accounted for in long-distance systems. The variance of the timing jitter can be expressed as

$$\langle \delta t^2 \rangle = t_0^2 \frac{2n_{sp}(G-1)^2}{9N_s G\ln(G)} \frac{L^3}{L_d^2 L_a} \tag{11}$$

where $N_s = 2P_s t_0/h\nu$ is the number of photons in the soliton. The fact that Gordon–Haus jitter grows cubically with distance makes it particularly important at transoceanic lengths, typically exceeding 1000 Km.

3.4.3. Acoustic Jitter. The electrostriction nonlinearity in the fiber gives rise to a mechanical pressure proportional to the optical intensity, which in turn modifies the refractive index of the fiber. In particular, an intense optical pulse like a soliton will give rise to a pressure (acoustic) wave moving radially outwards from the fiber center. The pulses in the wake of this wave will experience a randomly changing local refractive index, and hence (just as for Gordon–Haus jitter) a randomly changing carrier frequency that transforms into a timing jitter. An approximate expression for the standard deviation of this jitter has been found as [11]

$$\langle \delta t^2 \rangle^{1/2} = 0.0138 D^2 B L^2 (B - 1.18)^{1/2} \tag{12}$$

where D is the dispersion in ps/(nm·km), L is the system length in million meters [Mm], and B is the bit rate in gigabits per second (Gbps). Here a soliton separation $T/T_{FWHM} = 5$ has been assumed. We see that the variance of the jitter scales with distance to the fourth power: more rapid than Gordon–Haus jitter at large distances. This is because the acoustic perturbation is constant along the fiber, whereas the amplifier noise (which is the source for Gordon–Haus jitter) increases linearly with the system length.

3.5. WDM Considerations

Wavelength-division multiplexing (WDM) is another way of increasing the bit rate of soliton systems, where several frequency bands are used for solitons transmission. This technique was pioneered by Olsson et al. [12] 1991. The problem with WDM transmission using solitons stems mainly from collisions between solitons from different wavelength channels. In a perfectly ideal NLS equation solitons would collide elastically without changing carrier wavelength. However, the presence of losses and amplification may cause the collisions to be asymmetric if they occur around an amplifier, and a result will be a frequency displacement and a concomitant timing jitter. One solution to this problem is to force the collisions to be sufficiently long, forcing the collision distance to be larger than several amplifier spans. Widely separated WDM channels, however, collide over a short distance of fiber (simply because of dispersion) so this condition will restrict the accessible optical bandwidth to WDM solitons.

3.6. Soliton Control

Soliton control is the common name for methods to control the soliton parameters such as wavelength and position. There are two different approaches: passive and active control.

Passive soliton control has been suggested in the form of filters that are inserted along the transmission path. This helps to keep the soliton wavelength fixed. In this way not only Gordon–Haus and acoustic jitter can be remedied, but also interaction jitter and WDM-collision-induced jitter. A problem with this kind of filtering is that it defines a spectral region with excess gain, in which amplifier noise will grow excessively. A way around that problem is to slightly shift the center wavelength of the filters along the transmission path. In that way the solitons will follow the frequency shift, but the linear noise will not. Such sliding filter experiments have demonstrated 8×10 Gbps WDM soliton transmission over 10000 Km [13].

Active control usually acts in the time domain by using phase or amplitude modulators to retime and reshape the solitons. Using this technique, 10 Gbps over unlimited distances [14] has been demonstrated. However, this kind of active reshaping of the pulses suffers from the same drawbacks as the conventional electronic regeneration, specifically, incompatibility with WDM, complexity, and high cost.

3.7. Polarization Effects

A single-mode fiber is not strictly single-mode, since it always allows for two polarization modes. Imperfections along the fiber cause birefringence that will accumulate randomly during propagation, and eventually cause a net birefringence that is nonnegligible. This effect, known as *polarization-mode dispersion* (PMD), is considered by several researchers as a fundamental limit for linear pulse propagation since it drifts randomly as the fiber cable is exposed to temperature and/or pressure changes. Soliton pulses, however, are found to be more robust to PMD [15] than linear pulses. The reason for this is that an attractive interaction force between the polarization states works to keep the pulses together. However, some amount of radiation is shed by the solitons as a result of the random birefringence of the fiber, and that gives rise to pulse broadening, although not as significant as that for linear pulses. The soliton robustness to PMD was recently verified by transmission on installed fibers [16].

PMD will also limit the benefit of using polarization multiplexing to suppress pulse interactions. It has been shown [17] that for high PMD, a copolarized pulsetrain will perform better than a polarization multiplexed one.

There is another timing jitter effect stemming from the birefringence of the fiber, the so-called birefringence-mediated timing jitter [18]. This effect arises from the fact that amplifier noise will give rise to a small uncertainty in the soliton polarization state, which via the birefringence is transformed to a timing jitter. Also WDM collisions of soliton trains will give rise to polarization changes, which will add to this timing jitter. This source of jitter can be reduced with inline synchronous modulation. Passive filters, however, will not work in this case, as no frequency jitter is associated with the timing displacement.

4. DISPERSION-MANAGED SOLITONS

4.1. Introduction

Since the very late 1990s, solitons have become substantially more attractive through the rapidly emerging strategy of improving the performance of soliton transmission with dispersion management (DM). While the DM strategy, which involves altering the local dispersion between a large positive and a large negative GVD (group velocity dispersion) such that the average GVD is small, has long been used in linear systems it was only relatively recently appreciated that the same technique, if properly implemented, gives rise to several very striking improvements over conventional soliton transmission systems. While DM solitons are clearly nonlinear pulses, they are by no means classical solitons.

From a commercial viewpoint, however, the most important benefit with using DM solitons is that they, in principle, can use the already installed conventional fibers (with zero dispersion at 1300 nm) allowing a much more cost-effective upgrade together with dispersion-compensating fibers (DCFs) or chirped fiber gratings (then likely collocated with inline EDFAs), at certain intervals in the link.

4.2. Properties

DM solitons are not real solitons (they are rather solitary waves) in the sense that they have an inverse scattering transform that can be used to find the most relevant properties. Instead they emerged from extensive simulation work, and it was quite surprising to many researchers that the simulations revealed such stable and strictly periodic pulses.

In Fig. 1 the evolution of the temporal and spectral widths for a DM soliton is shown. Also shown is the dispersion map, which is a plot of how the dispersion changes with propagation distance. A lossless model is assumed, which is applicable to dense amplifier

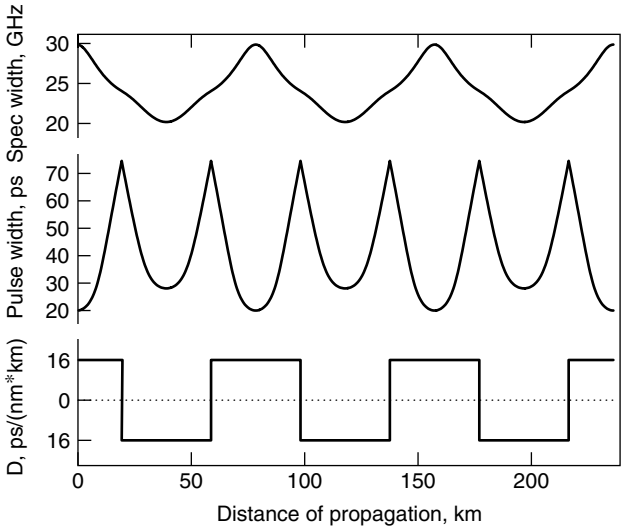


Figure 1. The evolution of the spectral width (top) temporal width (middle) of a DM soliton with zero average dispersion and map strength $S \approx 4$. The bottom plot shows the dispersion map. The symmetry of the evolution within the map period is due to the lossless model employed. (Figure contributed by A. Berntsson.)

spacing, but even when losses are included the qualitative conclusions and general properties will still hold. One difference in the lossy case is, however, that the symmetry of the propagation dynamics is removed, and the spectral dynamics will be concentrated to the positions in the dispersion map where the power is largest.

As a result of massive simulation work done by many groups the following properties of DM solitons have been found.

- The pulsewidth, chirp, and spectral width oscillates periodically in the dispersion map. There are two points within the period at which the pulse is chirp-free, and those correspond to local minima of the pulsewidth. One of those points is the global minimum width, referred to as the "shortest pulse" below.
- A central parameter that is useful for the characterization of DM solitons is the *map strength*, $S = (L_1|\beta_1''| + L_2|\beta_2''|)/T_{\mathrm{FWHM}}^2$, where L is length, β'' is dispersion, T_{FWHM} is the minimum pulsewidth in the full-width half-maximum sense, and subscripts 1,2 refer to the two fibers in the dispersion map. Physically the map strength is the number of dispersive lengths the pulse propagates in one period. DM solitons have been found for map strengths ranging from $S = 0$ (which is the same as conventional solitons) to $S \approx 12$, although this upper limit is a transition regime in which the pulses radiate and a perfect periodic evolution never arises.
- DM solitons have been found for anomalous, normal, and zero average dispersion $\overline{\beta}''$, defined as $\overline{\beta}'' = (L_1\beta_1'' + L_2\beta_2'')/(L_1 + L_2)$. Normal average dispersion is only possible for map strengths above 3.9.
- The shape of the shortest pulse ranges from hyperbolic secant at $S = 0$ to Gaussian for higher map strengths, and this is also evident from the time–bandwidth product, which increases with S from 0.32 (Sech) at $S = 0$ to 0.44 (Gaussian) and even higher for large values of S. In addition, the shortest pulse has oscillating tails in the pulse wings.
- The energy of a DM soliton pulse is enhanced relative to a soliton with the same average dispersion and pulsewidth.
- The interaction between DM solitons is less than that of conventional solitons, and an optimum map strength exists that minimizes the interaction.

The fact that DM solitons can work for an average net zero GVD and even for normal dispersion, was a striking and unexpected feature that would not work with conventional solitons. This can be understood by the fact that the spectral width of the pulses are larger when they propagate in the local anomalous dispersion regime than when they propagate in the normal dispersion regime (see Fig. 1). The pulses then effectively sense a net anomalous dispersion that is balanced by the nonlinearity as for conventional solitons.

The technical improvements with DM solitons over conventional ones are numerous. The signal-to-noise ratio is improved since the DM solitons have a larger peak power than do the corresponding conventional solitons. DM solitons have less Gordon–Haus and acoustic timing jitter, since the system average GVD is much smaller in these systems. A very important added benefit appears in wavelength-division-multiplexed (WDM) systems. Because of the alteration between large positive and negative GVD along the path, the jitter induced from WDM, soliton collisions is greatly reduced. This, in turn, allows for very dense WDM, which will improve the spectral efficiency substantially. The soliton PMD robustness is maintained or improved for DM solitons.

An important practical consequence of using DM solitons is that they reduce (or eliminate) the need for inline soliton control such as synchronous modulation or sliding filters. Yet soliton control methods are still applicable and will give improvement in terms of signal-to-noise ratio for DM solitons as well.

Quite impressive circulating loop experiments including WDM have been reported. In the study by Le Guen et al. [19] 51 very densely packed WDM channels each operating at 20 Gbps were transmitted over 1000 km with 100-km sections of standard fiber, clearly demonstrating the strength of the DM soliton technique.

4.3. Intrachannel Impairments

DM solitons have so many attractive features that they are likely to be implemented commercially in the near future. However, there are some novel transmission impairments, unique to DM solitons that need to be accounted for and analyzed in more detail. They are the so-called intrachannel effects; intrachannel four-wave mixing (ICFWM), and intrachannel cross-phase modulation (ICXPM) [20], and arise due to the nonlinear interaction between two neighboring pulses. Four-wave mixing (FWM) and cross-phase modulation (XPM) are

usually effects associated with WDM transmission. However, the fact that DM solitons are chirped and broadened, will cause different frequency components from neighboring pulses within the same wavelength channel to overlap in time, thereby causing FWM and/or XPM.

ICFWM arises for two neighboring pulses, that via four-wave mixing (FWM), creates new frequency components that in the time domain will give rise to a new pulse (commonly referred to as a "ghost pulse"), next to the two. The ghost pulse will then give rise to intersymbol interference and reduction of the eye opening. ICFWM is most prominent for large map strengths and high power.

ICXPM can be viewed as DM soliton interaction, and physically, it manifests as the frequency shift of one pulse induced by the presence of a neighboring pulse, which, by the dispersion transforms into a timing jitter. The effect can be minimized by selecting proper map strength and prechirp of the DM soliton.

As a rule, however, it seems that map strengths in the range 1–8 make best use of the unique features of DM solitons. This means that the use of installed standard fiber becomes difficult at very high bit rates (say, beyond 30–40 Gbps) as shorter pulses require a more rapidly varying dispersion map to maintain a proper S value. This is a limitation similar to the amplifier spacing limitation in conventional soliton systems, but it is much less severe. If an option is to install new dispersion-shifted fiber, on the other hand, this limitation becomes essentially unimportant.

5. EXPERIMENTS AND FIELD TRIALS

5.1. Soliton Pulse Sources

When doing soliton experiments, be it conventional or DM solitons, particular importance is placed on the properties of the pulse source, as it sets the lower limit of the system performance. A high-bit-rate soliton pulse source needs to produce low-chirp, low-timing-jitter pulses with proper duration (in the picosecond regime), repetition rate (10–40 GHz) and shape.

One possible choice is gain-switched (GS) laser diodes, possibly with an external cavity for tunability. However, they suffer the drawback of producing pulses that are strongly chirped, asymmetric and often too wide.

For laboratory experiments fiber ring lasers (FRLs) are very attractive, as they provide wavelength and pulse width tunability, besides meeting the above mentioned demands. Their drawback is that they are bulky, need active stabilization and sometimes also temperature control to achieve long-term stability.

Finally, it appears quite clear that electroabsorption modulators (EAMs) [whether integrated or not with a distributed feedback (DFB) laser] are very useful and simple sources for soliton transmission. While such sources were developed for linear NRZ (non-return-to-zero) systems, they have now proven to be near ideal in soliton systems as well. Although EAMs are not commercially available at 40 GHz yet, they are likely to be so within the near future.

Special considerations need to be taken in DM soliton systems, however, as the launch condition is different than for conventional solitons. The pulses should have a linear chirp such that it fits seamlessly in the periodically induced chirp variation along the link. This can be achieved by incorporating a proper length of fiber (or chirped fiber grating) in the transmitter once the overall dispersion map is known.

5.2. Loop Experiments

In order to investigate really long distances (megameters) of transmission, the loop experiments were developed in the early 1990s. This means that the data pulses are injected in a loop consisting of transmission fiber and amplifier, and then left to propagate a number of laps corresponding to a certain transmission distance. Acoustooptic switches are used to switch the pulse train in and out from the loop at proper time intervals. The drawback of loop experiments is that they may be poor models of reality when it comes to things like dispersion variation along the fiber, PMD or various kinds of drifts that may arise. In addition, a real system has more options to fine-tune, for example, amplifiers along the transmission line. However, as long as these drawbacks are recognized, loop experiments are very powerful indeed, and invaluable in lab evaluations of long-distance transmission.

5.3. Field Trials

In the field, many transmission link design restrictions and fiber properties make the systems far from optimal. The actual fiber parameters are nonperiodic with propagation distance as the systems are straight lines rather than relatively short loops. Both loss (in particular when including many contacts and splices along the link) and the PMD are typically much higher than in the laboratory. In particular, the PMD is higher since the fiber is no longer wound on a small drum making the mode coupling length longer, but also often simply because the installed fiber is old, being made before PMD was considered a real obstacle. The dispersion (or more specifically the zero-dispersion wavelength) might vary significantly along a fiber span. In addition, it may not be possible to tailor the dispersion map and amplifier locations to reach an optimal state. All these examples of nonidealities justify the need for field experiments.

Several soliton field experiments have been conducted in Japan by NTT [21–25], in the United States by MCI/Pirelli [26], and in Europe by the Advanced Communication Technologies & Services (ACTS) projects [27–30]. This is a good indication that solitons are indeed foreseen as very interesting candidates in commercial systems. Table 1 summarizes some data for the 10 soliton field experiments conducted from 1995 through 1999s. Much of the work has been at a bit rate of 40 Gbps, which is natural as this is expected to be the next standard trunk TDM rate. Again, it is not very easy to compare the results as the situation in each case differs. All the systems operated in the 1550 nm range, used optical time-division demultiplexing to the 10 Gbps electronic base rate, and the average loss/km ranged from 0.24 to 0.33 dB/km. Dispersion-shifted fiber was always used, apart from in

Table 1. Overview of the Soliton Field Experiments Conducted to Date[a]

Capacity (Gbps × 10^6 m)	Soliton Source	L_a (km)/ G (dB)	PMD (ps/km$^{1/2}$)/ DGD · T^{-1}	T/T_{FWHM}	Fiber	Ref./Year
10 × 2.7	EAM/GS	90/—	—	5	DSF	[25]/1995
10 × 2	FRL/GS	55/16	—	5	DSF	[21]/1995
10 × 0.3	—	50/12	0.04 / 0.7 %	2	SMF	[27]/1998
10 × 0.9	—	75/20	0.9 / 27 %	3.3	SMF+DCF	[26]/1998
4 × 10 × 0.45	—	75/20	0.9 / 27 %	3.3	SMF+DCF	[26]/1998
20 × 2	FRL	55/16	—	5	DSF	[22]/1995
40 × 1	FRL	55/16	selected	5	DSF/DCF	[24]/1998
40 × 1.4	FRL	55/16	selected	5	DSF/DCF	[25]/1998
40 × 0.4	FRL	57/19	0.3 / 24 %	2.5	DSF	[28]/1999
40 × 0.5	EAM	100/24	0.25 / 22 %	2.5	DSF	[29]/1999
80 × 0.2	FRL	57/19	0.11 / 12 %	2.8	DSF	[30]/1999

[a] Here, G and L_a denote the amplifier gain and separation; T, the bit separation; DGD, the differential group delay; and T/T_{FWHM}, the relative soliton separation.

two cases [26,27], where standard fiber was used. The study by Robinson et al. [26] deserves particular attention because DCF was used for dispersion compensation, which makes this the only DM soliton field experiment to date. In addition, that study [26] describes the transmission of four WDM channels (4 × 10 Gbps) but then over half the distance (450 km).

Polarization multiplexing was used in four experiments [27–30] and this serves mainly to allow the use of relatively wide pulses, which in turn allows for larger amplifier spans. Polarization multiplexing, however, is not as useful if the PMD of the system is high, as then the orthogonally polarized pulses would start to drift statistically in time relative to each other thereby creating intersymbol interference and increasing the soliton interaction. In the 40-Gbps cases and above, PMD was found to be the main capacity limiting factor. In two cases a special selection of low-PMD fiber was made [24,25].

Most of the more recent experiments used a mode-locked fiber ring laser (FRL) as a source, probably because these provide excellent pulse quality as well as tunability in terms of wavelength and pulse width. Other experiments used either gain-switched (GS) lasers or electroabsorbtion modulators (EAMs).

In only one case [25] was inline soliton control used (in the form of intensity modulation), and this experiment also achieved the highest capacity (54 Tbps · km).

Future soliton field trials are expected to (1) take advantage of the now well-known strategy of improving soliton transmission performance with dispersion management, as this method is very attractive for upgrading existing fiber plants; (2) implement dense WDM (in non-DS fiber lines) to boost aggregate capacity; (3) utilize different forms of inline control, particularly at high bit rates; and (4) further address the implications of PMD and techniques to combat it. The interesting WDM-TDM tradeoff for optimization of overall aggregate capacity will depend on the details of the fiber line parameters.

6. EVALUATION AND FUTURE OUTLOOK

To conclude, we note that the motivation for using solitons as information carriers have changed over the years. The property of being resistant to dispersive broadening was originally the main feature, but this was considered less important when the dispersion-compensating fibers became commercially available. Instead, this led to the development of the dispersion-managed soliton. The advantage of the DM soliton over linear transmission are features like the large power (which enables high signal-to-noise ratio) and PMD robustness. On the other hand, the difference between linear and nonlinear pulses are becoming increasingly fuzzy, and perhaps the distinction should be made between RZ (return-to-zero) and NRZ modulation rather than between linear and nonlinear transmission.

It is nevertheless interesting to note that solitons are now not only considered for oceanic systems but also for shorter terrestrial systems. There are still several challenges and opportunities remaining in order to take full advantage of solitons and to reach a better understanding. PMD remains an important topic that is not entirely understood when using solitons. Further work is also needed on WDM soliton and very-high-speed TDM soliton-systems. The use of DM solitons is a very recently established technique and there are thus many issues to consider. These include studies of robustness to deviations of optimum conditions, such as improper pulse launch condition, impact of nonperiodic dispersion maps and of PMD (both of which are difficult to study in loop experiments), and intrachannel effects. Nevertheless, solitons have now reached a level of maturity such that commercialization seems very near.

BIOGRAPHIES

Magnus Karlsson was born in Gislaved, Sweden, 1967. He received his M. Sc in engineering physics in 1991 and his Ph.D in electromagnetic field theory in 1994 from Chalmers University of Technology, Gothenburg, Sweden. The title of his Ph.D thesis is "Nonlinear propagation of optical pulses and beams." Since 1995, he has been with the Photonics Laboratory at Chalmers, first as assistant professor, and since 2001 as associate profesor in photonics. At the Photonics Lab his research has been devoted to fiber optic communication

systems, in particular transmission aspects such as fiber nonlinearities, solitons, four-wave mixing and polarization effects. He has authored or coauthored around 50 journal articles, 30 conference contributions, and two patents.

Peter A. Andrekson received his M.S. degree in electical engineering in 1984, and his Ph.D. degree in optoelectronics in 1988 from Chalmers University of Technology, Gothenburg, Sweden. During 1989–1992 he was with AT&T Bell Laboratories, Murray Hill, New Jerse, working on high speed fiber-optic transmission systems. He returned to Chalmers University in 1992 where he currently holds a professorship in photonics. Dr. Andrekson is the author and coauthor of over 200 technical publications and conference papers, and holds three patents. He also serves on several technical conference program committees. In 2000 he was awarded the Telenor Nordic Research Award for his contribution to optical technologies. Since 2000 he is on leave from Chalmers University, working as director of research at CENiX Inc. His interests cover essentially all aspects of high-capacity optical fiber communications.

BIBLIOGRAPHY

1. A. Hasegawa and F. Tappert, Transmission of stationary nonlinear optical pulses in dispersive dielectric fibers. I. Anomalous dispersion, *Appl. Phys. Lett.* **23**(3): 142–144 (1973).
2. L. F. Mollenauer, R. H. Stolen, and J. P. Gordon, Experimental observation of picosecond pulse narrowing and solitons in optical fibers, *Phys. Rev. Lett.* **45**(13): 1095–1098 (1980).
3. K. Iwatsuki, S. Nishi, M. Saruwatari, and M. Shimizu, 2.8 Gbit/s optical soliton transmission employing all laser diodes, *Electron. Lett.* **26**(1): 1–2 (1990).
4. V. E. Zakharov and A. B. Shabat, Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media, *Sov. Phys. JETP* **34**: 62–69 (1972).
5. J. Satsuma and N. Yajima, Initial value problems of one-dimensional self-modulation of nonlinear waves in dispersive media, *Progr. Theor. Phys. Suppl.* **55**: 284–306 (1974).
6. N. Akhmediev and M. Karlsson, Cherenkov radiation emitted by solitons in optical fibers, *Phys. Rev. A* **51**(3): 2602–2607 (1995).
7. P. K. A. Wai, H. H. Chen, and Y. C. Lee, Radiations by "solitons" at the zero group-dispersion wavelength of single-mode optical fibers, *Phys. Rev. A* **41**(1): 426–439 (1990).
8. K. J. Blow and N. J. Doran, The asymptotic dispersion of soliton pulses in lossy fibres, *Optics Commun.* **52**(5): 367–370 (1985).
9. J. P. Gordon, Interaction forces among solitons in optical fibers, *Opt. Lett.* **8**(11): 596–598 (1983).
10. J. P. Gordon and H. A. Haus, Random walk of coherently amplified solitons in optical fiber transmission, *Opt. Lett.* **11**(10): 665–667 (1986).
11. E. M. Dianov, A. V. Luchnikov, A. N. Pilipetskii, and A. M. Prokhorov, Long-range interaction of picosecond solitons through excitation of acoustic waves in optical fibers, *Appl. Phys. B* **B54**(2): 175–180 (1992).
12. N. A. Olsson et al., Bit-error-rate investigation of two-channel soliton propagation over more than 10,000 km, *Electron. Lett.* **27**(9): 695–697 (1991).
13. L. F. Mollenauer and P. V. Mamyshev, Massive wavelength-division multiplexing with solitons, *IEEE J. Quant. Electron.* **34**(11): 2089–2102 (1998).
14. M. Nakazawa et al., Experimental demonstration of soliton data transmission over unlimited distances with soliton control in time and frequency domains, *Electron. Lett.* **29**(9): 729–730 (1993).
15. L. F. Mollenauer, K. Smith, J. P. Gordon, and C. R. Menyuk, Resistance of solitons to the effects of polarization dispersion in optical fibers, *Opt. Lett.* **14**(21): 1219–1221 (1989).
16. B. Bakhshi et al., Experimental observation of soliton robustness to polarisation dispersion pulse broadening, *Electron. Lett.* **35**(1): 65–66 (1999).
17. X. Zhang, M. Karlsson, P. A. Andrekson, and E. Kolltveit, Polarization-division multiplexed solitons in optical fibers with polarization-mode dispersion, *IEEE Photon. Technol. Lett.* **10**(12): 1742–1744 (1998).
18. L. F. Mollenauer and J. P. Gordon, Birefringence-mediated timing jitter in soliton transmission, *Opt. Lett.* **19**(6): 375–377 (1994).
19. D. Le Guen et al., Narrow band 1.02 Tbit/s (51*20 Gbit/s) soliton DWDM transmission over 1000 km of standard fiber with 100 km amplifier spans, *Proc. OFC/IOOC'99. Optical Fiber Communication Conf. and Int. Conf. Integrated Optics and Optical Fiber Communications*, 1999, pp. PD4–1–PD4–3.
20. R. J. Essiambre, B. Mikkelsen, and G. Raybon, Intra-channel cross-phase modulation and four-wave mixing in high-speed TDM systems, *Electron. Lett.* **35**(18): 1576–1578 (1999).
21. M. Nakazawa et al., Field demonstration of soliton transmission at 10 Gbit/s over 2000 km in Tokyo metropolitan optical loop network, *Electron. Lett.* **31**(12): 992–994 (1995).
22. M. Nakazawa et al., Soliton transmission at 20 Gbit/s over 2000 km in Tokyo metropolitan optical network, *Electron. Lett.* **31**(17): 1478–1479 (1995).
23. K. Iwatsuki et al., Field demonstration of 10 Gb/s-2700 km soliton transmission through commercial submarine optical amplifier system with distributed fiber dispersion and 90 km amplifier spacing, *Proc. 21st European Conf. Optical Communication, ECOC'95*, 1995, pp. 987–990.
24. A. Sahara et al., Single channel 40 Gbit/s soliton transmission field experiment over 1000 km in Tokyo metropolitan optical loop network using dispersion compensation, *Electron. Lett.* **34**(22): 2154–2155 (1998).
25. K. Suzuki et al., 40 Gbit/s soliton transmission field experiment over 1360 km using inline soliton control, *Electron. Lett.* **34**(22): 2143–2145 (1998).
26 N. Robinson et al., 4xSONET OC-192 field installed dispersion-managed soliton system over 450 km of standard fiber in the 1550 nm Erbium band, *Proc. Optical Fiber Communication Conf.* 1998, pp. PD19–1–PD19–4.

27. P. Franco et al., 10 Gbit/s alternate polarisation soliton transmission over 300 km step-index fibre link with no inline control, *Electron. Lett.* **34**(11): 1116–1117 (1998).

28. E. Kolltveit et al., Single-wavelength 40 Gbit/s soliton field transmission experiment over 400 km of installed fibre, *Electron. Lett.* **35**(1): 75–76 (1999).

29. F. Matera et al., Impact of polarisation mode dispersion in field demonstration of 40 Gbit/s soliton transmission over 500 km, *Electron. Lett.* **35**(5): 407–408 (1999).

30. J. Hansryd et al., 80 Gbit/s single wavelength soliton transmission over 172 km installed fibre, *Electron. Lett.* **35**(4): 313–315 (1999).

FURTHER READING

Agrawal G. P., *Nonlinear Fiber Optics*, 2nd ed., Academic Press, San Diego, 1995.

Hasegawa A. and Y. Kodama, *Solitons in Optical Communications*, Clarendon Press, Oxford, 1995.

Mollenauer L. F., J. P. Gordon, and P. V. Mamyshev, Solitons in high bit rate long-distance transmission, in I. P Kaminow and T. L. Koch (eds.), *Optical Fiber Telecommunications III A*, Academic Press, San Diego, 1997.

Taylor J. R. (ed.), *Optical Solitons-Theory and Experiment*, Cambridge Univ. Press, Cambridge, UK, 1992.