Nanoelectronics and Photonics

Nanostructure Science and Technology

Series Editor: David J. Lockwood, FRSC
National Research Council of Canada
Ottawa, Ontario, Canada

Current volumes in this series:

Functional Nanostructures: Processing, Characterization and Applications
Edited by Sudipta Seal

Light Scattering and Nanoscale Surface Roughness
Edited by Alexei A. Maradudin

Nanotechnology for Electronic Materials and Devices
Edited by Anatoli Korkin, Evgeni Gusev, and Jan K. Labanowski

Nanotechnology in Catalysis, Volume 3
Edited by Bing Zhou, Scott Han, Robert Raja, and Gabor A. Somorjai

Nanostructured Coatings
Edited by Albano Cavaleiro and Jeff T. De Hosson

Self-Organized Nanoscale Materials
Edited by Motonari Adachi and David J. Lockwood

Controlled Synthesis of Nanoparticles in Microheterogeneous Systems
Vincenzo Turco Liveri

Nanoscale Assembly Techniques
Edited by Wilhelm T.S. Huck

Ordered Porous Nanostructures and Applications
Edited by Ralf B. Wehrspohn

Surface Effects in Magnetic Nanoparticles
Dino Fiorani

Interfacial Nanochemistry: Molecular Science and Engineering at Liquid-Liquid Interfaces
Edited by Hitoshi Watarai

Nanoscale Structure and Assembly at Solid-Fluid Interfaces
Edited by Xiang Yang Liu and James J. De Yoreo

Introduction to Nanoscale Science and Technology
Edited by Massimiliano Di Ventra, Stephane Evoy, and James R. Heflin Jr.

Alternative Lithography: Unleashing the Potentials of Nanotechnology
Edited by Clivia M. Sotomayor Torres

Semiconductor Nanocrystals: From Basic Principles to Applications
Edited by Alexander L. Efros, David J. Lockwood, and Leonid Tsybeskov

Nanotechnology in Catalysis, Volumes 1 and 2
Edited by Bing Zhou, Sophie Hermans, and Gabor A. Somorjai

(Continued after index)

Anatoli Korkin · Federico Rosei
Editors

Nanoelectronics and Photonics

From Atoms to Materials, Devices, and Architectures

Editors
Anatoli Korkin
Nano and Giga Solutions
Gilbert, AZ
USA
korkin@nanoandgiga.com

Federico Rosei
Initiative National de la Recherche Scientifique, Énergie, Matériaux et Télécommunications
Université du Québec
Québec, QC Canada
rosei@emt.inrs.ca

Series Editor
David J. Lockwood, FRSC
National Research Council of Canada
Ottawa, Ontario, Canada

ISBN: 978-0-387-76498-6 e-ISBN: 978-0-387-76499-3
DOI: 10.1007/978-0-387-76499-3

Library of Congress Control Number: 2008931856

© 2008 Springer Science+Business Media, LLC
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science + Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Cover illustrations: 1. Fragment of an atomic scale model of Mo-HfO_2 interface (Chapter 7); 2. Photoluminescence spectra computed for different times after the 1s excitation with classical field (Chapter 10); 3. Electron distribution across the nanowire, for the wire width of 30 nm (*left panel*) and 8 nm (*right panel*) (Chapter 6); 4. A *field-programmable nanowire interconnect* (FPNI) structure (Chapter 4); 5. Modulated nanoindentation of a ZnO nanobelt with an atomic force microscope tip (Chapter 9); 6. Ferromagnet/antiferromagnet bilayers (Chapter 5); 7. A woodpile structure of a 3D photonic crystal (Chapter 11); 8. A *scanning electron microscope* (SEM) image of a structure fabricated *by two-photon polymerization* (2PP) technique, which resembles pulmonary alveoli – microcapillaries responsible for gas exchange in the mammalian lungs (Chapter 12); 9. The time evolution of the space charge region during *deep level transient spectroscopy* (DLTS) measurements (Chapter 8).

Printed on acid-free paper

springer.com

Preface

Tutorial lectures given by world-renowned researchers have become one of the important traditions of the first two days of the *Nano and Giga Challenges* (NGC) conference series. Soon after preparations for the first forum in Moscow, Russia, had begun, the organizers realized that publication of the lectures notes from NGC2002 would be a valuable legacy of the meeting and a significant educational resource and knowledge base for students, young researchers, and experts alike. Our first book was published by *Elsevier* and received the same title as the meeting itself – *Nano and Giga Challenges in Microelectronics* [1]. Our second book, *Nanotechnology for Electronic Materials and Devices* [2] based on the tutorial lectures at NGC2004 in Krakow, Poland, and the current book from NGC2007 in Phoenix, Arizona, have been published in Springer's *Nanostructure Science and Technology* series.

Nanotechnology as the art (i.e., science and technique) of control, manipulation, and fabrication of devices with structural and functional attributes smaller than 100 nm (0.1 μm) is perfectly suited to advanced CMOS technology. This technology holds the capacity for massive production of high-quality nanodevices with an enormous variety of applications from computers to biosensors, from cell phone to space shuttles, and from large display screens to small electronic toys.

Exponential growth of the number of transistors in commercial integrated circuits (ICs) was first identified as a trend in 1965 by G. Moore, Intel's co-founder. Later recognized as *Moore's law*,[1] this trend has become an imperative and, until recently, almost a religious prophecy as documented in the International Technology Roadmap for Semiconductors (ITRS).[2] However, scaling of transistors and other devices to smaller and smaller sizes, which has provided the basis for this exponential growth, has limits, physical (size of the atoms), technological (lithography) and economic (see articles of K. Likharev and S. Williams), which will be

[1] The number of transistors that can be placed on a commercial integrated circuit is increasing exponentially, doubling approximately every 2 years: G.E. Moore, Electronics, vol. 38, No. 8, 1965.

[2] http://www.itrs.net/

reached by CMOS technology in the next decade. The exponential growth will converge into an S-curve, a well-known trend in biology and economics.

Will this pessimistic forecast result in decreasing interest in society (and in funding!) for electronics research? Is any feasible alternative to CMOS technology available in the near future from photonics, molecular electronics, or revolutionary engineering solutions, such as departure from two-dimensional ICs on the surface of silicon wafers to three-dimensional structures? All these *gigantic challenges* and potential *nanotechnology solutions* are actively debated at *Nano & Giga Forums*. We invite you to *google* the date and location of our next meeting and join us in learning, active discussion, information exchange, and networking in the vibrant and dynamic atmosphere of next NGC forum!

The success of the NGC2007 conference in Arizona, which resulted in the publication of this book and in other contributions making up special issues of *Nanotechnology*[3] and *Solid State Electronics*,[4] would have not been possible without generous support from many sponsors and research institutions. We gratefully acknowledge the contributions and support of Arizona State University (conference host and co-organizer), International Science and Technology Center (ISTC), National Science Foundation (NSF), Defense Advanced Research Agency (DARPA), Office of Naval Research, Army Research Office, Computational Chemistry List (CCL), Springer, City of Tempe, STMicroelectronics, Quarles & Brady LLP, Oak Ridge National Lab, Canadian Consulate in Phoenix, Salt River Project (SRP), and many other local, national and international, and individual supporters.

Special thanks to Ms. Megha Chadha, graduate student from Singapore University, for careful reading of the book chapters and other assistance with NGC2007 publications.

Anatoli Korkin
Co-founder of Nano & Giga Forum
and president of Nano and Giga Solutions, Inc.

References

1. J. Greer, A. Korkin, J. Labanowski (eds) *Nano and Giga Challenges in Microelectronics* (Elsevier, Amsterdam, Netherlands, 2003).
2. A. Korkin, E. Gusev, J. Labanowski, S. Luryi (eds) *Nanotechnology for Electronic Materials and Devices*, (Springer, New York, 2007).

[3] Selected and invited papers from NGC2007 symposium on nanoCMOS technology (guest editors S. Goodnick, A. Korkin, T. Naito, and N. Peyghambarian) published in Solid State Electronics, vol. 51, No. 10, 2007.

[4] Selected and invited papers from NGC2007 symposium on molecular and biolectronics (guest editors P. Krstic, E. Forzani, NJ Tao, and A. Korkin) published in Nanotechnology, vol. 18, No. 42, 2007.

Tutorial Lectures from Nano & Giga Forum 2007

Federico Rosei

It has been a great honor and pleasure for me to serve as one of the Guest Editors for this volume of tutorial lectures from the latest edition of the 'Nano and Giga' conference. Participating in the last two meetings (2004 in Poland and more recently 2007 in Arizona), and in particular co-editing this tutorial book, has been an exciting and rewarding experience and has significantly broadened my scientific horizons.

This book contains useful chapters that can be used as reference and lecture material for advanced undergraduate and graduate courses. Each tutorial is a useful, self-contained lecture written for non-experts and the contents of this volume cover a broad range of research topic at the forefront and state of the art.

My personal fascination with 'nanoscience' relates to the new (i.e., different from the bulk form) properties that a material may exhibit when at least one of its dimensions is reduced below 100 nm. I have always been fascinated by the Periodic Table of the Elements, and frustrated at the same time: why are there only 92 stable elements? Nanoscience partly resolves this frustration: since each element behaves differently (often in surprising ways) at the nanoscale, it gives the opportunity to extend, so to speak, the Periodic Table introducing new dimensions to it.

Today 'Nano', a prefix widely used in modern science (from the Greek word for dwarf), is an intrinsically rich and multidisciplinary field of research, as it represents a natural convergence of disciplines [1]. As such, it provides an excellent opportunity for scientific education in a broad sense, going back to Galileo and Newton, the founders and fathers of modern science. From a fundamental point of view, 'Nano' has given us a new understanding of materials and their properties, namely how many characteristics may change dramatically at small scales due to an increased surface-to-volume ratio or to quantum effects or to a combination of factors. Examples of this include the new allotropes of carbon (carbon nanotubes, fullerenes and more recently graphene), as

Professor, INRS-MT, University of Quebec Canada Research Chair in Nanostructured Organic and Inorganic Materials

well as the appearance of luminescence properties in nano-silicon [2], and finally the different optical, chemical and electronic properties that gold exhibits at the nanoscale [3].

In terms of harnessing such properties into useful applications, Nanotechnology holds the promise of addressing the great challenges of humanity in the 21st century, namely the access to clean and renewable energies, preserving and protecting the environment and improving human health. It is my hope that more and more time and resources will be devoted to developing 'nanoresearch' in these specific areas, as these are the ones that are more likely to have a positive and beneficial impact on our society as a whole.

Nowadays fewer and fewer scientists are willing to take the time to write a good book chapter. In today's world, dominated by impact factors and citation indices, service to the community (in the form of teaching or writing a chapter to be used as lecture material) is unfortunately undervalued. Under these circumstances I am particularly grateful to all the authors of the chapters contained in this volume for doing an overall excellent job and for honoring their initial commitment.

We hope you enjoy these pages and find them useful to further your education or for your research. If you have suggestions for future Nano and Giga tutorial series of specific topics not addressed here, please do not hesitate to let us know as the readership's feedback and advice is our only way to gauge how we can improve.

References

1. G.A. Horley, 'The Importance of Being "Nano"', *Small* **2**, 3 (2006).
2. L.T. Canham, *Appl. Phys. Lett.* **57**, 1046 (1990).
3. A. Sugunan, J. Dutta, 'Nanoparticles for Nanotechnology', *PSI Jilid* **4**, 50 (2004).

Contents

Part I Perspectives

1 Nanotechnology: A Scientific Melting Pot 3
Nicolaas Bloembergen

2 Integrated Circuits Beyond CMOS 5
Konstantin K. Likharev

3 Nano and Giga Challenges for Information Technology 9
R. Stanley Williams

Part II Tutorial Lectures

4 Hybrid Semiconductor-Molecular Integrated Circuits for Digital Electronics: CMOL Approach 15
Dmitri B. Strukov

5 Fundamentals of Spintronics in Metal and Semiconductor Systems 59
Roland K. Kawakami, Kathleen McCreary, and Yan Li

6 Transport in Nanostructures 115
Stephen M. Goodnick

7 Density Functional Theory of High-*k* Dielectric Gate Stacks 171
Alexander A. Demkov

8 Trapping Phenomena in Nanocrystalline Semiconductors 191
Magdalena Lidia Ciurea

9 Nanomechanics: Fundamentals and Application in NEMS Technology 223
Marcel Lucas, Tai De Li, and Elisa Riedo

10 Classical and Quantum Optics of Semiconductor Nanostructures . . . 255
Walter Hoyer, Mackillo Kira, and Stephan W. Koch

11 Photonic Crystals: Physics, Fabrication, and Devices 353
Wei Jiang and Michelle L. Povinelli

12 Two-Photon Polymerization – High Resolution 3D Laser Technology and Its Applications. 427
Aleksandr Ovsianikov and Boris N. Chichkov

Index 447

Contributors

Nicolaas Bloembergen
University of Arizona, nbloembergen@optics.arizona.edu

Boris N. Chichkov
Laser Zentrum Hannover e.V., b.chichkov@lzh.de

Magdalena Lidia Ciurea
National Institute of Materials Physics, ciurea@infim.ro

Alexander A. Demkov
The University of Texas at Austin, demkov@physics.utexas.edu

Stephen M. Goodnick
Arizona State University, Stephen-Goodnick@asu.edu

Walter Hoyer
Philipps-University Marburg, Walter.Hoyer@physik.uni-marburg.de

Wei Jiang
Rutgers University, wjiangnj@ece.rutgers.edu

Roland K. Kawakami
University of California, roland.kawakami@ucr.edu

Mackillo Kira
Philipps-University Marburg

Stephan W. Koch
Philipps-University Marburg, stephan.w.koch@physik.uni-marburg.de

Tai-De Li
Georgia Institute of Technology

Yan Li
yan.li002@email.ucr.edu

Konstantin K. Likharev
Stony Brook University, klikharev@notes.cc.sunysb.edu

Marcel Lucas
Georgia Institute of Technology, marcel.lucas@gatech.edu

Kathleen McCreary
kathleen.mccreary@email.ucr.edu

Aleksandr Ovsianikov
Laser Zentrum Hannover e.V., A.Ovsianikov@lzh.de

Michelle L. Povinelli
Stanford University, mpovinel@stanford.edu

Elisa Riedo
Georgia Institute of Technology, elisa.riedo@physics.gatech.edu

Dmitri B. Strukov
Hewlett-Packard Laboratories, dmitri.strukov@hp.com

R. Stanley Williams
Hewlett-Packard Laboratories, stan-williams@hp.com

Part I
Perspectives

Chapter 1
Nanotechnology: A Scientific Melting Pot

Nicolaas Bloembergen

When a linear dimension of a device or a theoretical subject of investigation is smaller than 1 μm, it may be said that a one-dimensional nanoregime has been entered. In this sense the study of monomolecular and bimolecular layers and surface physics in general is now said to belong to nanoscience. More recently the study of surfaces has been enhanced by the techniques of nonlinear optical spectroscopy, by scanning tunneling spectroscopy and by atomic force microscopy.

The ancient use of submicron colloidal particles of gold and silver in glass to obtain colored window materials is an early example of three-dimensional nanotechnology. It is based on the range of plasmon-resonant frequencies in small metallic particles.

A small number of atomic layers of GaAs and GaAlAs or other semiconducting compounds have led to light-emitting diodes and lasers over a wide frequency range. Such layered structures have also created two-dimensional plasmas of conduction electrons which exhibit quantum Hall effects. Small semiconducting particles called quantum dots may function as versatile submicroscopic light sources.

Biological and medical investigations have also focused increasingly on nanostructures during the past two decades. Genetics and neurophysiology are concerned with the detailed structure of individual molecules, including DNA, RNA and various enzymes and proteins on cell walls or other substrates.

Material scientists have found the structure of new carbon molecules, including the buckeyball C_{60} and other Buckminster fullerenes. Carbon fibers are very strong and highly conducting nanomaterials. The drive in computer technology to ever smaller dimensions, evidenced by Moore's law, has led not only to electronic transistors and switches with nanodimensions, but also to very small optical devices, including lasers and nonlinear optical couplers.

Since 2001 the US federal budget has included the National Nanotechnology Institute. This NNI has played a key role in fostering cross-disciplinary

N. Bloembergen
College of Optical Sciences, University of Arizona, Tucson, AZ, USA
e-mail: nbloembergen@optics.arizona.edu

A. Korkin, F. Rosei (eds.), *Nanoelectronics and Photonics*,
DOI: 10.1007/978-0-387-76499-3_1, © Springer Science+Business Media, LLC 2008

networks and partnerships. Several universities have built new laboratories to serve as Nanoscale Sciences and Engineering Centers. They provide a home for faculty from many departments, including physics, chemistry, biology, material science, computer science, neuroscience, genetics and others.

This interdisciplinary effort has an impact on the traditional academic organization of strictly autonomous academic departments in separate scientific disciplines. It encourages the establishment of interdisciplinary formal courses, both at the undergraduate and graduate levels. In industrial research organizations this intermingling of disciplines has always been more common in order to reach a well-defined technical goal.

My education was strictly as a physicist, but after my formal studies I have always enjoyed my contracts with other disciplines. My research interests in magnetic resonance, lasers and nonlinear optics have provided ample opportunities to interact with chemists, biologists and medical doctors. Because of advancing age and medical doctors I have not actively participated in the recent trend toward nanotechnology. Therefore my introductory lecture deals not with nanometer spatial dimensions, but with very small temporal dimensions. It is remarkable that the duration of laser pulses has been shortened by 15 orders of magnitudes in four decades. My lecture, entitled "From millisecond to attosecond laser pulses," reviews the historical developments toward ever smaller time scales. They are mostly based on diverse nonlinear optical phenomena. The text of my remarks has been published in Progress in Optics 50, 1–12, 2007, edited by E. Wolf since its inception in 1961.

I apologize for my tangential connection with nanotechnology. This field will undoubtedly continue to contribute to further progress in optics, as well as to many other disciplines, since it is truly a scientific melting pot.

Chapter 2
Integrated Circuits Beyond CMOS

Konstantin K. Likhareu

Semiconductor microelectronics, based on silicon CMOS circuits, is arguably the most successful technology ever developed by mankind because it sustained its fast, exponential (*Moore's Law*) progress for several decades. As a result, this technology has become the basis of all current information technology revolution. However, now scientists and engineers agree that this progress will run into what is called the *red brick wall* of physical, technical, and economical limitations some time during the next decade. Optimists believe this crisis may be deferred until the 22-nm ITRS technology node, to be reached by 2015 or so, while the pessimists like myself do not see any realistic way for the technology to go beyond the 32-nm node, to be reached by 2013 or maybe even a year or two earlier. In any case, the range of opinions (of well-informed professionals) is rather narrow, and continues to shrink.

The negative impact of running into the red brick wall for the high-tech economy may be hardly exaggerated. Sure, whatever happens after that point, there will be more and more silicon chips fabricated each year. However, if the exponential progress of the key metrics, most notably the circuit cost per unit device, has been stopped or slowed down to a crawl, the integrated circuit manufacturing, as virtually all mature manufacturing industries, will most probably be outsourced to countries with cheaper labor. The current electronics industry giants, which currently live on innovation, will face a survival challenge. This is why the extension of Moore's Law into the sub-10-nm range is such a vital task. As usual, there are both good and bad news from the current battle on this *nanoelectronic* frontier.

On the positive side, both the federal government and electronic industry leaders now recognize the necessity and urgency of research in this direction. On the negative side, the efficiency of those efforts is very much questionable. Large electronic companies, being extremely efficient at moving up an evolutionary path such as semiconductor microelectronics, have serious problems with adapting revolutionary (*disruptive*) technologies like nanoelectronics. As a

K.K. Likhareu
Stony Brook University, New York, USA
e-mail: klikhareu@notes.cc.sunysb.edu

A. Korkin, F. Rosei (eds.), *Nanoelectronics and Photonics*,
DOI: 10.1007/978-0-387-76499-3_2, © Springer Science+Business Media, LLC 2008

result, the substantial resources thrown onto the problem by the companies, some states, and federal government (within the $1B/year-scale National Nanotechnology Initiative) are not, in my humble opinion, being spent effectively. Most of this money goes to groups studying various nanoscale objects (carbon nanotubes, semiconductor nanowires, DNA molecules, you name it) with little or no attempt to understand how exactly these objects would work as electron devices, and how these devices might be incorporated into an integrated circuit. It comes without saying that at such approach the vital questions about the possible fabrication cost and performance of future nanoelectronic circuits may not be even asked, leave alone answered.

Fortunately, the past year evidenced the emergence of a more systematic approach to nanoelectronics by a few (for now, just few) academic and industrial groups. Such approach naturally starts with the determination of the main reasons for the anticipated crisis. In contrast to what some industry captains declare, it is certainly the exponentially growing fabrication tool cost, dominated by that of circuit patterning equipment. Indeed, the workhorse device of CMOS circuits, the silicon MOSFET, requires an accurate lithographic definition of several dimensions including the length and width of its conducting channel. As these devices key are scaled down, arising quantum mechanical effects require the definition to be much more precise, which in turn requires much more expensive lithography tools. At some point, the scaling will start bringing diminishing returns. (The reason why this situation is not evident to everybody in the electronics industry is that the major chipmakers had outsourced the development of better patterning techniques to the fabrication equipment producers long ago, and right now those companies are probably not very interested in revealing the real, rather gloomy situation with tool progress to their customers.)

Another necessary component of the systemic approach to the microelectronics is a candid estimate of nanoelectronic devices. Unfortunately, such evaluations show that the nanodevices comparable in their functionality to silicon MOSFETs either run into similar fabrication problems, or cannot be assembled into integrated circuits, or both. The much-heralded *bottom-up* approach (e.g., device self-assembly) also has not given any encouraging results yet.

Fortunately, among all this doom and gloom there is a glim of hope. During the past several years, several groups, including our Stony Brook team, have simplified the decade-old idea of hybrid CMOS/nanoelectronic circuits in which the CMOS stack is augmented with a back-end nanoelectronic add-on. Most recent work in this field is focused on nanowire crossbar add-ons, with simple bistable two-terminal devices formed at each crosspoint, and area-distributed CMOS/nano interfaces – see, e.g., the detailed review article by D. B. Strukov, and a brief write-up by R. S. Williams in this collection, and references therein.

The basic idea of such hybrid circuits is to combine the advantages of CMOS technology (including its flexibility and high fabrication yield) with the enormous density of simple (two-terminal) nanodevices which may be fabricated

reproducibly, at reasonable cost, and naturally incorporated into the nanowire crossbar fabric. However, the main motivation for the hybrid circuit concept is that the nanowire crossbars, including the crosspoint devices, may be fabricated using advanced patterning techniques (such as nanoimprint, EUV interference lithography, block-copolymer lithography), while removing from these techniques the requirement of precise layer alignment. It is believed that the removal of this burden may enable, within the next 15–20 years, an improvement of the resolution of these techniques down to a few nanometers.

Recent detailed simulations have shown that the hybrid circuits with such fine features (though employing much larger MOSFETs fabricated using the ordinary photolithography) may provide at least a two-orders-of-magnitude advantage over purely CMOS ICs in such basic metrics as memory density, logic delay-by-area product, and image processing speed, at manageable power density and high defect tolerance. This leading edge is equivalent to the extension of the Moore's Law progress of microelectronics by approximately 10–15 years beyond the "red brick wall".

Simulations have also shown that the hybrid circuits may be used for operations in the mixed-signal mode as bio-inspired neuromorphic networks ("CrossNets") which can be used for performing several important information processing tasks (such as online recognition of a particular person in a large crowd) much more efficiently than digital circuits implementing the same algorithm. Moreover, estimates show that in the long run, CMOL CrossNets may challenge human cortical circuitry in density, far exceeding it in speed, at realistic power. Of course, in order to map these advantages on performing really intelligent information processing tasks, much work has to be carried out by interdisciplinary teams of theoretical neurobiologists, computer scientists, and electrical and computer engineers, but the possible technological and societal impact of such development may hardly be overestimated.

Of course, it may happen that other approaches to nanoelectronics will prove to be more fruitful than the hybrid circuit concept. However, I am confident that only the systemic approach to the problem, taking into account all its aspects, may lead us to success. Let me hope that this collection will be an important step in this direction.

Chapter 3
Nano and Giga Challenges for Information Technology

R. Stanley Williams

The primary technology driver for the integrated circuit industry and all of the information technology supported by that industry has been Moore's law, the observation that the number of transistors on a chip has roughly doubled every 18 months over the past four decades. In concert with this exponential increase in transistors has come the dramatic increase in performance of integrated circuits while the cost of a single chip has remained fairly constant. This astounding improvement in a basic technology over a many-decade-long period is unprecedented and has led to a huge industry with a major economic footprint and enabled major increases in productivity and functionality for a wide variety of other sectors of society.

There have been many eras in the past when pundits have predicted the end of Moore's scaling for a variety of excellent technical and engineering reasons. In all those cases, motivated engineers have overcome the barriers foreseen by the experts and kept the industry on the path to fulfilling the promise of more transistors for less money. However, in the twenty-first century we are quickly running up against a very fundamental obstacle, the granularity of matter. We will not be able to build device components with sizes that are a fraction of a single atom. Thus, we know that there is an end to "traditional" scaling of transistor sizes, but we cannot predict exactly when that will occur. This is because it is not just a physics or engineering issue, but also an economic question. If we invest enough money in a system, we can eventually achieve the ultimate performance that the laws of nature will allow, but the cost of doing so may be much larger than any possible return on that investment. Thus, the quest to continue *functional scaling*, e.g., the continued increase in *performance* of integrated circuits at *fixed cost per chip*, is a scientific and engineering exercise that is constrained by economics.

Perhaps the greatest challenges facing the integrated circuit industry as we approach fundamental limits are manufacturability, reliability and resiliency.

R.S. Williams
Director, Information and Quantum Systems Laboratory, HP Labs, Hewlett-Packard Company, 1501 Page Mill Road, MS 1123, Palo Alto, CA 94304
e-mail: stan.williams@hp.com

A. Korkin, F. Rosei (eds.), *Nanoelectronics and Photonics*,
DOI: 10.1007/978-0-387-76499-3_3, © Springer Science+Business Media, LLC 2008

Today's logic chips are perfect and they operate for very long times without experiencing an unplanned interruption. However, as the feature sizes of devices scale down to the few nanometer scale, the properties of the devices will vary more broadly because of the inevitable statistical fluctuations in the number of atoms in a component of a transistor or a wire. Since the circuits will be manufactured at a temperature above absolute zero, these fluctuations are assured by the Second Law of Thermodynamics. If the fluctuations are severe enough, the device will not work at all. We can thus see three major problems – devices in a circuit that are nonfunctional because of manufacturing errors, which we call defects; devices that yield incorrect results because of a fluctuation in a property during operation of the circuit, which we call faults; and devices that start out working properly but then experience a catastrophic event, which we call device death. Today, devices with any mistakes made during manufacture must be discarded, which has a negative impact on the manufacturing yield and increases the cost of the chips that are perfect. There are ways to handle faults during operation today, but if a single device on a chip dies while it is in service, the entire chip must be replaced. Thus, the nano and giga challenges for integrated circuits are that the probability of a problem with an individual component in a circuit is increasing dramatically with decreasing size and even worse the probability of failure of the system is increasing with the number of components on a circuit. With exponential scaling, we will very quickly cross the threshold from high-yield circuits that perform reliably for long periods to low-yield circuits that experience frequent interruptions and device deaths. The brute force way of dealing with imperfections caused by atomic-scale statistical fluctuations is to spend a lot of money improving the manufacturing processes. However, this approach can rapidly spiral out of control to make chip manufacturing too expensive to improve.

A more useful approach is to look at the fundamental architecture of a chip to see if it is possible to program in defect, fault and death tolerance. After all, it is well known that a substantial number of brain cells die every day; yet rather than fall over dead when the first brain cell dies, humans continue to operate for many decades and in at least the best of cases experience only a gradual (or hopefully graceful) degradation in capacity. In fact, this question of building reliable machines out of unreliable parts was a significant area of research by such giants of computer and information science as von Neumann [1] and Shannon [2] in the 1950 s. Although these early researches were interesting and informative, the entire area of thought was for the most part abandoned in the 1960 s when high-yielding and reliable transistors in integrated circuits came into being. It is only now that we are entering into the nano and giga age that we need to reexamine the issues of how to build reliable machines given that they will be manufactured with defects and experience faults and device deaths.

The major approaches to making a logic circuit reliable and resilient in the presence of defects, faults and deaths deal with optimization of redundancy of circuit elements. In some circuit architectures, one can plan to overprovision the system with extra components and the wiring to connect them into a circuit to

compensate for failed devices. This approach is known as reconfiguration [3] and requires that the entire circuit be analyzed to locate and catalog all of the defects. A computer program, for example a compiler, can then download a program onto the system and route around the broken components. This approach is extremely robust and can compensate for a significant percentage (~3%) of defective components in a system, but it has the disadvantage that it does not work for faults and the entire search and avoid strategy, which may be costly in terms of execution time, must be rerun periodically to deal with device deaths. Recently, Strukov and Likharev have proposed a variant of this type of scheme that combines CMOS components with nanoscale wires and switches to create a hybrid circuit (CMOL) [4] that can significantly improve the performance and defect tolerance of a *field-programmable gate array*, a type of logic circuit. Snider and Williams have proposed a variant of this architecture that may be significantly easier to manufacture but can still offer significant performance advantages [5] over CMOS-only circuits.

Another approach to building more robust circuits is to use coding theory to design and build redundant circuits that contain efficient automatic correction for errors of various sorts. Such an approach is effective for defects, faults and deaths, although it is limited in the types of functions that can be protected. There is no requirement to find the defective components – the existing devices will automatically compensate for any broken devices as long as the number of broken devices does not exceed the maximum number allowed by the code used. For example, a demultiplexer, the bridging unit that provides an interface between some level of CMOS driving circuitry with just a few devices and any explicitly nanoscale circuits with a large number of devices, can be made significantly more robust by the appropriate inclusion of extra data lines and devices [6]. A small amount of redundancy can provide an exponential increase in the reliability of a circuit, which is excellent in terms of keeping the cost of the error correction to a minimum. Given a particular known device failure rate and the desired level of reliability for the entire circuit, it is a straightforward (although certainly nontrivial) matter to identify a code, or geometric circuit layout, that will satisfy the constraints of the problem for certain types of operations (with a demultiplexer being the best example for efficiency). However, at this stage it does not appear possible to apply coding theory to general types of logic circuits.

It is possible, although not at all certain, that by combining reconfiguration and coding, one may be able to construct extremely resilient systems that defend against all error types. This is an active field of research. The primary problem is that by adding enough redundancy to fulfill both types of reliability enhancement, one may pay such a large circuit area penalty that it just makes more sense to stop scaling to smaller feature sizes and stay with a larger and thus more robust generation of CMOS. There are also other possibilities for new architectures, such as "neuromorphic computing," which utilizes synthetic synapses to perform a type of analog computation. There are certainly interesting times ahead as the approaches described here and possibly many others that have

not yet been invented are tried out and compared to strictly scaled CMOS in terms of cost, performance and reliability. This architectural work greatly complements materials and processing research, which seeks to improve the functionality and reliability of individual devices. Indeed, there is no stable ground, since there are continual new advances in both materials and information sciences that make it ever more likely that functional scaling of integrated circuits will continue for many more decades into the future. The primary issue is for both communities to keep in contact so that each can leverage the advances of the other.

References

1. J. von Neumann, “Probabilistic logics and the synthesis of reliable organisms from unreliable components” in C. E. Shannon and J. McCarthy, Eds. *Automata Studies* (1955), 43–98.
2. E. F. Moore and C. E. Shannon, “Reliable circuits using less reliable relays,” *Journal of the Franklin Institute* (1956), 191–208 and 281–297.
3. J. R. Heath, P. J. Kuekes, G. S. Snider and R. S. Williams, “A defect-tolerant computer architecture: Opportunities for nanotechnology,” *Science* **280** (1998), 1716.
4. D. B. Strukov and K. K. Likharev, “CMOL FPGA: A cell-based, reconfigurable architecture for hybrid digital circuits using two-terminal nanodevices,” *Nanotechnology* **16** (2005), 888–900.
5. G. S. Snider and R. S. Williams, “Nano/CMOS architectures using field-programmable nanowire interconnect,” *Nanotechnology* **18** (2007), art. no. 035204.
6. P. J. Kuekes, W. Robinett, G. Seroussi and R. S. Williams, “Defect-tolerant interconnect to nanoelectronic circuits: Internally redundant demultiplexers based on error-correcting codes,” *Nanotechnology* **16** (2005), 869–882.

Part II
Tutorial Lectures

Chapter 4
Hybrid Semiconductor-Molecular Integrated Circuits for Digital Electronics: CMOL Approach

Dmitri B. Strukov

Abstract This chapter describes architectures of digital circuits including memories, general-purpose, and application-specific reconfigurable Boolean logic circuits for the prospective hybrid CMOS/nanowire/nanodevice ("CMOL") technology. The basic idea of CMOL circuits is to combine the advantages of CMOS technology (including its flexibility and high fabrication yield) with those of molecular-scale nanodevices. Two-terminal nanodevices would be naturally incorporated into nanowire crossbar fabric, enabling very high function density at acceptable fabrication costs. In order to overcome the CMOS/nanodevice interface problem, in CMOL circuits the interface is provided by sharp-tipped pins that are distributed all over the circuit area, on top of the CMOS stack. We show that CMOL memories with a nano/CMOS pitch ratio close to 10 may be far superior to the densest semiconductor memories by providing, e.g., 1 Tbit/cm^2 density even for the plausible defect fraction of 2%. Even greater defect tolerance (more than 20% for 99% circuit yield) can be achieved in both types of programmable Boolean logic CMOL circuits. In such circuits, two-terminal nanodevices provide programmable diode functionality for logic circuit operation, and allow circuit mapping and reconfiguration around defective nanodevices, while CMOS subsystem is used for signal restoration and latching. Using custom-developed design automation tools we have successfully mapped on reconfigurable general-purpose logic fabric ("CMOL FPGA") the well-known Toronto 20 benchmark circuits and estimated their performance. The results have shown that, in addition to high defect tolerance, CMOL FPGA circuits may have extremely high density (more than two orders of magnitude higher that that of usual CMOS FPGA with the same CMOS design rules) while operating at higher speed at acceptable power consumption. Finally, our estimates indicate that reconfigurable application-specific ("CMOL DSP") circuits may increase the speed of low-level image processing tasks by more than two orders of magnitude as compared to the fastest CMOS DSP chips implemented with the same CMOS design rules at the same area and power consumption.

D.B. Strukov
Hewlett-Packard Laboratories, 1501 Page Mill Road, Palo Alto, CA 94304, USA
e-mail: dmitri.strukov@hp.com

A. Korkin, F. Rosei (eds.), *Nanoelectronics and Photonics*,
DOI: 10.1007/978-0-387-76499-3_4, © Springer Science+Business Media, LLC 2008

4.1 Introduction

The prospects to continue the Moore Law with current VLSI paradigm, based on a combination of lithographic patterning, CMOS circuits, and Boolean logic, beyond the 10 nm frontier are uncertain [1, 2]. The main reason is that at gate length beyond 10 nm, the sensitivity of parameters (most importantly, the voltage threshold) of MOSFETs to inevitable fabrication spreads grows exponentially. As a result, the gate length should be controlled with a few-angstrom accuracy, far beyond even the long-term projections of the semiconductor industry [3]. For example, for the most promising double gate silicon-on-insulator (SOI) MOSFETs, the definition accuracy of 5 nm long gate channel should be better than 0.2 nm in order to keep fluctuations of the voltage threshold below a reasonable value of 50 mV [1], i.e., much smaller than ITRS projected value of 0.5 nm [3]. Even if such accuracy could be technically implemented using sophisticated patterning technologies, this would send the fabrication facilities costs (growing exponentially even now) skyrocketing and lead to the end of the Moore's Law some time during the next decade.

Similar problems with scaling await existing memory technologies when their feature sizes will approach the 10 nm scale regime. Indeed, the basic cell (holding one bit of information) of today's mainstream memories, like static and dynamic random access memories, as well as those of relatively new but already commercialized technologies like ferroelectric, magnetic, and structural phase transition memories, needs at least one transistor and hence will run into the aforementioned limitation in the future.

Needless to say that the stoppage of Moore Law will have biggest consequences not only for semiconductor industry but also for computing society. Indeed, in addition to high-performance systems, e.g., supercomputers, which directly profit from faster and denser memory and logic circuits, there are plenty of emerging applications, such as image processing [4], which would greatly benefit from CMOS technology scaling. For example, the first step in hyperspectral imaging [5] for a realistic 12-bit 1024×1024 pixel array with 200 spectral bands requires a processing throughput of $\sim 10^{14}$ operations per second (100 Tops) and an aggregate data bandwidth of $\sim 10^{11}$ bits per second (100 Gbps) [6]. Even aggressively scaled hypothetical 22 nm multi-core Cell processor [7], which has been specifically designed for image processing tasks, falls far short of the prospective needs [8].

The main alternative nanodevice concept, single electronics [1, 9], offers some potential advantages over CMOS, including a broader choice of possible materials. Unfortunately, for room-temperature operation, the minimum features of these devices (single-electron islands) should be below ~ 1 nm [9]. Since the relative accuracy of their definition has to be between 10 and 20%, the absolute fabrication accuracy should be of the order of 0.1 nm, again far too small for the current and realistically envisioned lithographic techniques.

Fortunately, critical dimensions of devices can be controlled much more accurately via some other techniques, e.g., film deposition. Even more attractive would be a “bottom-up” approach, with the smallest active devices formed in a special way ensuring their fundamental reproducibility. The most straightforward example of such device is a specially designed and chemically synthesized molecule, implementing single-electron transistor.

However, integrated circuits consisting of molecular devices alone are hardly viable because of limited device functionality. Most importantly this is because the voltage gain of a 1 nm scale transistor, based on any known physical effect, can hardly exceed one,[1] i.e., the level necessary for sustaining the operation of virtually any active digital circuit. This is why the most plausible way toward high-performance nanoelectronic circuits is to integrate nanodevices, and the connecting nanowires, with CMOS circuits whose (relatively large) field-effect transistors would provide the necessary additional functionality, in particular high voltage gain.

The novel hybrid technology paradigm will certainly require rethinking of the current circuit architectures, which is exactly the focus of this review. First, we start with reviewing nanoscale devices suitable for such hybrid circuits (Section 4.2). The main challenges in prospective hybrid circuits and the effective solution offered by “CMOL” concept and its cousins will be outlined next (Section 4.3). In the rest of this chapter, we review our approach for CMOL-based digital memories (Section 4.4), general-purpose reconfigurable Boolean logic circuits (Section 4.5), and application-specific reconfigurable Boolean logic circuits (Section 4.6). Finally, in Section 4.7, we briefly summarize the results of our discussion.

4.2 Devices

The first critical issue in the development of semiconductor/nanodevice hybrids is making a proper choice in the trade-off between nanodevice simplicity and functionality. On the one hand, simple molecule-based nanodevices (like the octanedithiols [11]), which may provide nonlinear but monotonic $I-V$ curves with no hysteresis, are hardly sufficient for highly functional integrated circuits. Indeed, bistability of nanodevices helps to deal with regularity and defect tolerance of hybrid circuits – see Section 4.3. On the other hand, very complex molecular devices (like a long DNA strand [12]) may have numerous configurations that can be, as a matter of principle, used for information storage. However, such molecules are typically very “soft”, so that thermal fluctuations at room temperature (that is probably the only option for broad electronics

[1] For example, for the most prospective ballistic field-effect transistors, this is mainly due to leakage tunneling of thermally excited electrons. In single-electron transistors, the gain is limited by island to gate capacitance ratio. The gain of interference transistors is also typically small, see, e.g., Ref. [10].

applications) may lead to uncontrollable switches between their internal states, making reliable information storage and usage difficult, if not totally impossible.

Moreover, so far there are only practical solutions for fabricating two-terminal devices because they may have just one critical dimension (distance between the electrodes) which may be readily controlled by, e.g., film deposition or oxidation rate. Equally, chemically directed self-assembly of two-terminal devices would be immeasurably simpler than the multi-terminal ones. This is why many realistic proposals of hybrid circuits are based on two-terminal "latching switches" or "programmable diodes" (see, e.g., Refs. [13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25], as well as circuits described in this chapter [8, 26, 27, 28, 29, 30], and also recent reviews [31, 32, 33, 34, 35, 36, 37]).[2] The functionality of such devices is illustrated in Fig. 4.1a. At low applied voltages, the device behaves as a usual diode, but a higher voltage may switch it between low-resistive (ON) and high-resistive (OFF) states.

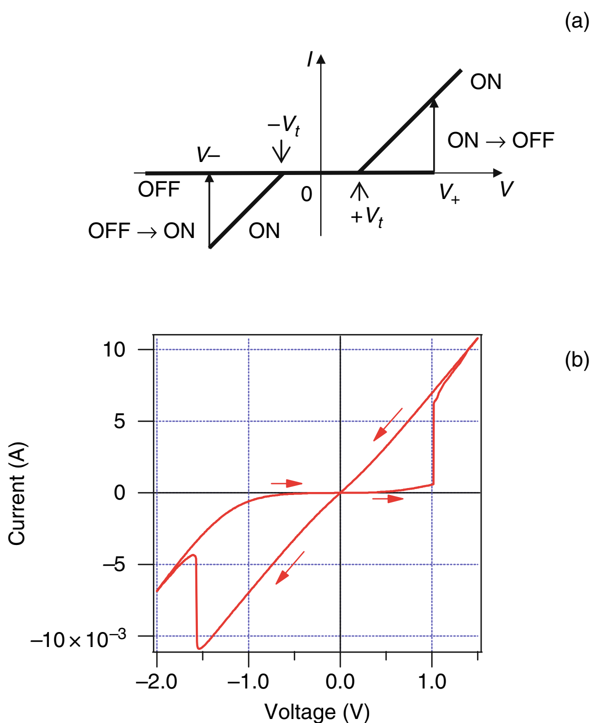


Fig. 4.1 $I - V$ curve of (**a**) two-terminal latching switch considered for this chapter (schematically) and (**b**) typical bipolar Pt–TiO_2–Pt resistive switch [46]

[2] As it will be shown later in this work, the diode-like characteristic is necessary for the operation of the hybrid memory circuits and is helpful for the proposed logic circuits. However, simple programmable resistance switches (Fig. 4.1b) could be enough for, e.g., nanoelectronic neuromorphic networks [38, 39, 40, 41], programmable interconnect hybrid CMOS/nanodevice architectures [42, 43], as well as Goto-pair-based circuit architectures [22, 44, 45]. The latter two concepts will be briefly discussed below.

Interestingly, the devices with a similar functionality based on amorphous oxides (typically Al, Si, Nb, and Ta) and chalcogenide glasses have been demonstrated almost half century ago, see, e.g., a very comprehensive review in Ref. [47]; however, neither of these device technologies was broadly accepted by electronic industry in the context of (random access) memory and logic circuit applications. Recently, bistable switching was demonstrated for much broader choice of material systems which can be crudely organized in the following categories[3]:

- Relatively thick organic films, both with [52, 53, 54, 55] and without [56, 57, 58, 59, 60] embedded metallic clusters
- Self-assembled monolayers (SAM) of molecules [61, 62, 63, 64]
- Thin chalcogenide glass layers [65, 66, 67, 68, 69, 70]
- Semiconductor films [71, 72]
- Amorphous or polycrystalline (nonstoichiometric) oxides, e.g., SiO and AlO [73], with most notable group involving transition metal oxides, such as TiO_2 [46, 74, 75, 76, 77], Nb_2O_5 [78], CuO [79], NiO [80, 81, 82], CoO [81, 83], VO_2 [84, 85], and various perovskite oxides [86, 87, 88, 89, 90, 91, 92, 93]

Despite tremendous surge of research activity in thin-film switches it is still too early to claim success. The most common problems are reproducibility of *I–V*s from device to device, large variations of set/reset threshold voltage (or current), and shifts of characteristics upon repeated cycling. In fact, even probing whether there are any fundamental problems with scaling in such devices is precluded by poor understanding of physics of the ON–OFF switching.

Indeed, the microscopic nature of resistance switching and charge transport is still under debate in both organic and inorganic structures [47, 49, 51, 87]. For example, perovskite structures exhibit very diverse electrical properties, and hence switching models based on ferroelectricity [93], magnetism [94], and metal–insulator [84, 88, 91] transitions have been proposed. Alternatively, bistability due to electron charge trapping for either defect-rich crystalline or amorphous oxides which modulates the impurity band conduction was speculated [82, 90]. Even though the electronic band gap is quite high for most of the oxides, one cannot exclude transport trough conduction band also. This is why several mechanisms based on Schottky barrier modulation either, via trapping of electrons on the interface or due to band bending were also investigated [89].

It is worth noting that many ingenious experiments have been devised to elucidate the nature of switching – see, e.g., Refs. [77, 87, 95, 96, 97]. On the other hand, understanding of experimental results is very often complicated by the profusion of different behaviors observed in nanoscale switches (i.e., bipolar vs. unipolar switching, ohmic vs. non linear I–Vs with or without negative

[3] For more extensive review of thin-film devices, see, e.g., Refs. [47, 48, 49, 50, 51].

differential slope, smooth or sharp threshold ON–OFF switching) which are not always fully reported in literature.

The lack of good physical model precludes further optimizations of device structure and most importantly screening less promising candidates and focusing on the most prospective ones. For example, in nonhomogeneous or filamentary conduction, the transport is due to some random active conducting centers such as hopping percolation paths, separated by distances of the order of a few nanometers. In order to be reproducible, the device should have a large number of such centers. This is why the extension of the excellent reproducibility demonstrated for such statistical devices with a lateral size larger than 100 nm [79] to the most interesting range, i.e., below 10 nm, might present a challenge. On the other hand, homogenous switching, e.g., due to drift of oxygen vacancies inside the oxide film [98] would suffer less from such limitation of the law of large numbers. In fact, a few percent nonstochiometric oxide may have hundreds of oxygen vacancies (dopants) in $\sim$100 nm^3 volume.

Even better prospects might hold uniform self-assembled monolayers of specially designed molecules [38] implementing binary single-electron latching switches [99]. A major challenge for molecular devices is the reproducibility of the interface between the monolayer and the second (top) metallic electrode, because of the trend of the metallic atoms to diffuse inside the layer with molecules during the electrode deposition [100], and the difficulty in ensuring a unique position of the molecule relative to the electrodes, and hence a unique structure and transport properties of molecular-to-electrode interfaces. Very encouraging proposal toward solution of these problems is to include relatively large "floating electrodes" as shown in Fig. 4.2 [32]. If the characteristic internal resistance R_0 of such a molecule is much higher than the range of possible values

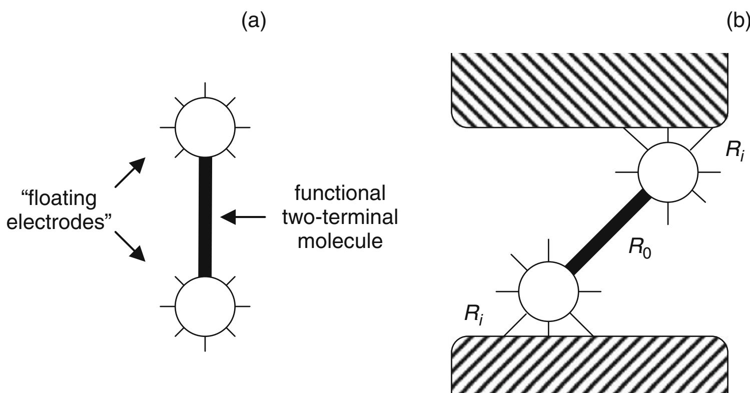


Fig. 4.2 A molecule with "floating electrodes" (**a**) before and (**b**) after its self-assembly on "real electrodes", e.g., metallic nanowires (schematically) [32]

of molecule/electrode resistances R_i, and the floating electrode capacitances are much higher than those of the internal single-electron islands, then the transport through the system will be determined by R_0 and hence be reproducible. Another possible way toward high yield is to form a self-assembled monolayer (SAM) on the surface of the lower nanowire level, and only then deposit and pattern the top layer (with the option of inserting a conductive polymer interlayer between SAM and the metal electrode). Such approach has already given rather reproducible results (in the nanopore geometry) for simple, short molecules [11, 101].

Finally, the potentially enormous density of nanodevices can hardly be used without individual contacts to each of them. This is why the fabrication of wires with nanometer-scale cross-section is another central problem of nanoelectronics. The currently available photolithography methods, and even their rationally envisioned extensions, will hardly be able to provide such resolution. Several alternative techniques, like the direct e-beam writing and scanning-probe manipulation, can provide a nanometer-scale resolution, but their throughput is forbiddingly low for VLSI fabrication. Self-growing nanometer-scale-wide structures like carbon nanotubes or semiconductor nanowires can hardly be used to solve the wiring problem, mostly because these structures (in contrast to the nanodevices that have been discussed above) do not have means for reliable placement on the lower integrated circuit layers with the necessary (a few nm) accuracy. Alternatively, in principle, vertically stacked semiconductor nanowires might be used to build $\sim 5\,\text{nm} \times 5\,\text{nm}$ area transistors [102, 103]. However, it is unclear whether the yield of such epitaxially vertically grown nanowires can be high enough for large-scale integration. Even more importantly, interconnecting such dense array of vertically stacked nanowires presents a challenge unless macroscale CMOS wires are used.

Fortunately, there are several new patterning methods, notably nanoimprint [104, 105, 106], block-copolymer technology [107], and interference lithography [108, 109], which may provide much higher resolution than the standard photolithography. Indeed, the layers of parallel nanowires with a nano half-pitch $F_{\text{nano}} = 17\,\text{nm}$ have already been demonstrated [110], and there are good prospects for the half-pitch reduction to 3 nm or so in the next decade [104, 105, 106]. (The scaling of the pitch below 3 nm value would be not practical because of the quantum mechanical tunneling between nanowires.)

4.3 Circuits

The novel device and patterning technologies may allow to extend microelectronics into the few-nanometer range. However, they impose a number of challenges and limitation for integrated circuit design.

• **Defect tolerance** – Perhaps, the main challenge faced by the hybrid circuits might be the requirement of very high defect tolerance. Indeed, it is natural to

expect that at the initial stage of development of all nanodevices, their fabrication yield for $F_{nano} < 30$ nm will be considerably below 100%, and, for $F_{nano} \sim 3$ nm, will possibly never approach this limit closer than a few percent. This number can be compared with at most 10^{-8}% of bad transistors for the mature CMOS technology [3].

It is somewhat believed that the most numerous and hence the most significant types of "hard" (fabrication-induced) faults will be "stuck-on-open" defects in nanodevices. Such defects correspond to permanently disconnected crosspoints. Typically, it is assumed that stuck-on-open defects are uniformly distributed with probability q. (Note that any clustering of defects would be much easier to cope with via reconfiguration – see, e.g., next section.) This assumption is justified by recent experimental works [111]. It is important, therefore, for an architecture to provide first of all the defect tolerance with respect to these kind faults. This is why only these kinds of defects were taken into account in most of the hybrid circuit papers [8, 27, 28, 30, 42, 112, 113]. Among other types of defects in hybrid circuits the most significant are broken/shorten nanowires and "stuck-on-close" defects, corresponding to permanently connected crosspoints. Typically, such defects are much harder to tolerate, e.g., see defect tolerance analysis in Refs. [18, 26, 26, 114]. This is because in the most realistic scenario bad nanowires (for "stuck-on-close" defects it is those nanowires which are connected to a given defective crosspoint) together with all potentially good nanodevices connected to these nanowires should be excluded.

- **Circuit regularity** – Nanoimprint and interference lithography cannot be used for the fabrication of arbitrary integrated circuits, in particular because they lack adequate layer alignment accuracy ("overlay"). This means that the nanowire layers should not require precise alignment with each other. The remedy to this problem can be a very regular "crossbar" nanowire structure [115] with two layers of similar wires perpendicular to those of the other layers (Fig. 4.3). On the one hand, such structures are ideal for the integration of two-terminal nanodevices which can be sandwiched, e.g., by self-assembly or film deposition, in between two layers of nanowires. On the other hand, if all nanodevices are functionally similar to each other, the relative position of one nanowire layer with respect to the other is not important. Not surprisingly,

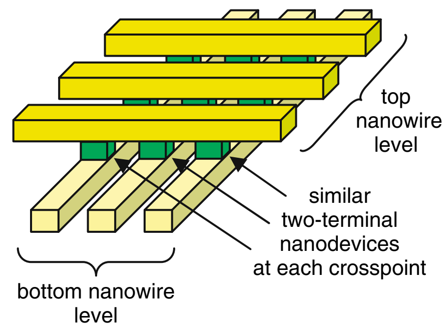


Fig. 4.3 Crossbar array structure

virtually all proposals for digital CMOS/nanodevice hybrids, most importantly including memories [18, 19, 27, 30, 64, 111, 116, 117] and Boolean logic circuits [8, 13, 14, 15, 16, 17, 20, 21, 22, 23, 24, 25, 26, 28, 29, 42, 44, 45, 113, 118, 119], are based on crossbar structures (see also reviews of such circuits in Refs. [31, 32, 33, 34, 35, 36, 37]).[4]

Naturally, it is the regularity of crossbar structures that necessitates bistability in nanodevices. The specific functionality of crossbar-based logic circuits is achieved with configuration of nanodevices (essentially via disabling some devices by programming them in the OFF state and leaving active devices in the ON state – see more detailed discussion in Section 4.5).

• **Micro-to-nano interface** – The lack of alignment accuracy of novel patterning technologies also results in much harder problem of building CMOS-to-nanowire interfaces. In fact, the interface should enable the CMOS subsystem, with a relatively crude device pitch $2\beta F_{\mathrm{CMOS}}$ (where $\beta \sim 1$ is the ratio of the CMOS cell size to the wiring period and F_{CMOS} is a CMOS half-pitch), to address each wire separated from the next neighbors by a much smaller distance F_{nano}.

Several solutions to this problem, which had been suggested earlier, seem to be not very efficient. In particular, almost all of the proposed interfaces are based on statistical formation of semiconductor-nanowire field-effect transistors gated by CMOS wires [120, 121, 122, 123] and can only provide a limited (address decoding-type) connectivity, which might present a problem for sustaining sufficient data flow in and out of the nanoscale subsystem. Moreover, such demux-based interfaces present architectural challenges since they are both needed for configuration of the nanodevices, as well as for transferring data between CMOS and nano subsystems. Also, the technology of ordering chemically synthesized semiconductor nanowires into highly ordered parallel arrays has not been developed, and there is probably no any promising idea that may allow such assembly.

A more interesting approach was discussed in Ref. [16] (see also Refs. [33] and [124]). It is based on a cut of the ends of nanowires of a parallel-wire array, along a line that forms a small angle $\alpha = \arctan(F_{\mathrm{nano}}/F_{\mathrm{CMOS}})$ with the wire direction. As a result of the cut, the ends of adjacent nanowires stick out by distances (along the wire direction) differing by $2F_{\mathrm{CMOS}}$ and may be contacted individually by the similarly cut CMOS wires. Unfortunately, the latter (CMOS) cut has to be precisely aligned with the former (nanowire) one, and it is not clear from Ref. [16] how exactly such a feat might be accomplished using available patterning techniques.

Figure 4.4 shows the so-called CMOL approach [1, 32, 125] to the interface problem. The difference between this approach (based on earlier work on the so-called InBar neuromorphic networks [38, 39]) and the suggestions discussed

---

[4] Another, not less exciting, application of the crossbar nanoelectronic hybrids, neuromorphic networks [38, 39, 40, 41], is out of the scope of this work.

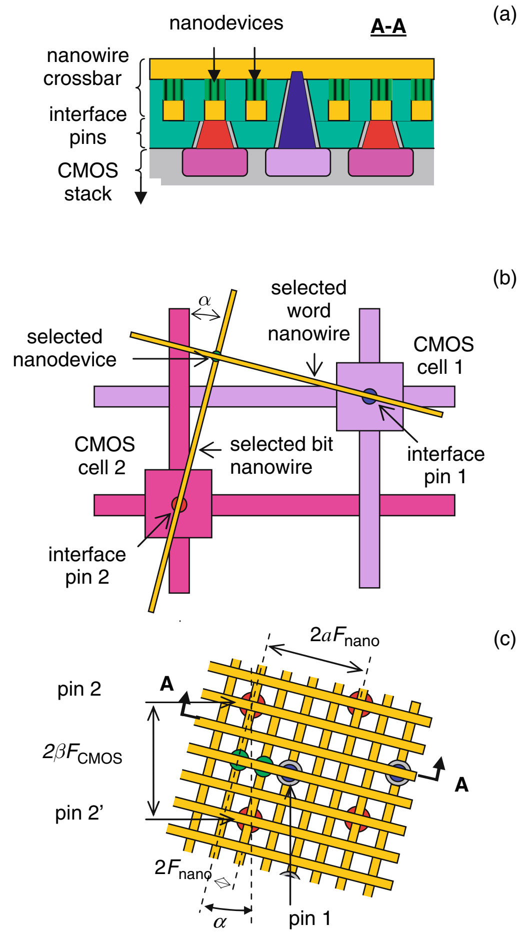


Fig. 4.4 The generic CMOL circuit: (**a**) a schematic side view, (**b**) a schematic top view showing the idea of addressing a particular nanodevice via a pair of CMOS cells and interface pins, and (**c**) a zoom-in top view on the circuit near several adjacent interface pins. On panel (**b**), only the activated CMOS lines and nanowires are shown, while panel (**c**) shows only two devices. (In reality, similar nanodevices are formed at all nanowire crosspoints.) Also disguised on panel (**c**) are CMOS cells and wiring (*See* Color Insert)

above is that in CMOL the CMOS-to-nanowire interface is provided by pins distributed all over the circuit area. In the generic CMOL circuit (Fig. 4.4), pins of each type (contacting the bottom and top nanowire levels) are located on a square lattice of period $2\beta F_{\mathrm{CMOS}}$. Relative to these arrays, the nanowire crossbar is turned by a (typically, small) angle α which is found as (Fig. 4.4c)

$$\alpha = \arctan\frac{1}{a} = \arcsin\frac{F_{\mathrm{nano}}}{\beta F_{\mathrm{CMOS}}} \ll 1, \tag{4.1}$$

where a is a (typically, large) integer. Such tilt ensures that a shift by one nanowire (e.g., from the second wire from the left to the third one in Fig. 4.4c) corresponds to the shift from one interface pin to the next one

(in the next row of similar pins), while a shift by a nanowires leads to the next pin in the same row. This trick enables individual addressing of each nanowire even at $F_{\mathrm{nano}} \ll \beta F_{\mathrm{CMOS}}$. For example, the selection of CMOS cells 1 and 2 (Fig. 4.4c) enables contacts to the nanowires leading to the left one of the two nanodevices shown on that panel. (The simplest circuitry enabling such selection would be CMOS pass transistor – see Section 4.4 for more discussion of this point.) Now, if we keep selecting cell 1, and instead of cell 2 select cell 2' (using the next CMOS wiring row), we contact the nanowires going to the right nanodevice instead.

It is also clear that a shift of the nanowire/nanodevice subsystem by one nanowiring pitch with respect to the CMOS base does not affect the circuit properties. Moreover, a straightforward analysis of CMOL interface (Fig. 4.5) shows that at an optimal shape of the interface pins (for example, when top radius of both upper and lower level interface pins, the nanowire width, and nanowire spacing are all equal) even a complete lack of alignment of these two subsystems leads to a theoretical interface yield of 100%. (Note that the last statement is only true for the latest version of CMOL [26, 29] in which pin, going to the upper nanowire level, intentionally interrupts a lower layer wire – see Fig. 4.4.) Even if the interface yield will be less than 100%, it may be acceptable, taking into account that the cost of the nanosystem fabrication, including the chemically directed assembly of molecular devices, may be rather low, especially in the context of an unparalleled density of active devices in CMOL circuits.

More recently, several approaches to the interface between CMOS and nano subsystems, very similar to CMOL, have been proposed. In Ref. [126], interface between nano and CMOS wires is supposed to be formed by exposing portions of CMOS wires with precisely angled cut in the insulator layer (Fig. 4.6). The key point in this proposal is that the interface yield can be up to 100% without any overlay alignment between nano and CMOS layers if the vertical gap w_{gap} between CMOS openings and its height is exactly equal to nanowire width w_{nano} and nanowire spacing s_{space}, correspondingly. Clearly, the idea behind it is the same as that of CMOL, if one replaces CMOS area openings with CMOS pillars.

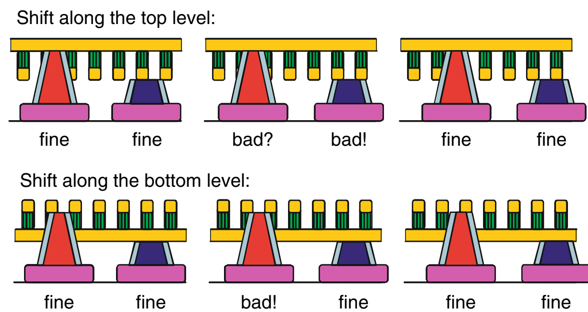


Fig. 4.5 The idea of 100% CMOS-to-nano interface yield without any overlay alignment

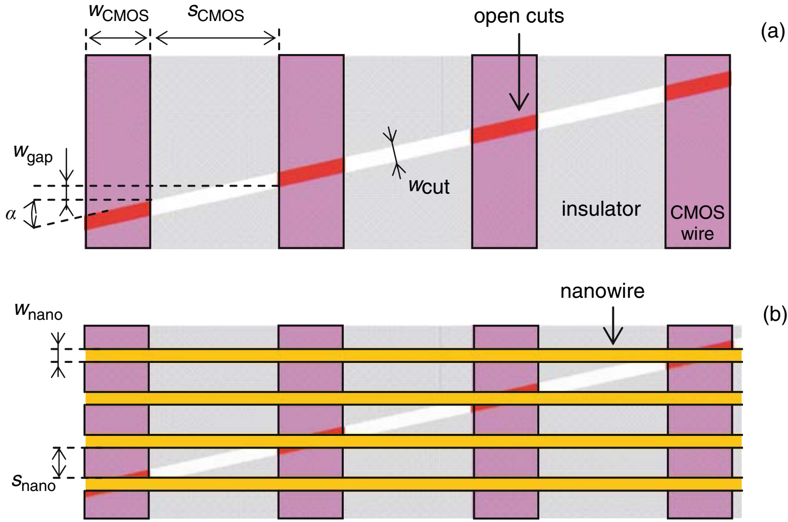


Fig. 4.6 Peripheral CMOS-to-nano interface [126]

The advantage over CMOL approach is that the cut is much easier to implement than the pins. On the other hand, this approach has also rather substantial disadvantages: (I) The interface density is more than twice lower than that of the maximum possible one; (II) the proposed interface is peripheral since the suggested technique is only feasible for interfacing one layer of nanowires at a time. Hence, it may be used on the crossbar periphery rather than distributed over all the area as CMOL. As a result, the implementation of logic circuits in this technology is hardly feasible (cf. Section 4.5).

The area interface without nanometer-scale pins is suggested recently in HP's FPNI circuits [42]. According to the authors, such FPNI circuits are a generalization of the CMOL FPGA approach, allowing for simpler fabrication and more conservative process parameters. More specifically, authors indicate that the sharply pointed interface pins with nanometer-scale top radii present a fabrication challenge and at the initial stage it is easier to replace them with CMOS-scale pins. For such change, the nanowire crossbar requires CMOS-scale alignment with respect to CMOS subsystem and will be much sparser than the original used in CMOL (Fig. 4.7). Another feature that simplifies fabrication of FPNI is the fact that nanodevices are used only as programmable resistance switches. The downside of FPNI approach is that more functionality is transferred in CMOS subsystem and together with sparser nanowire crossbar the areal density of FPNI logic circuits is substantially lower than that of CMOL-based ones [42]. The performance

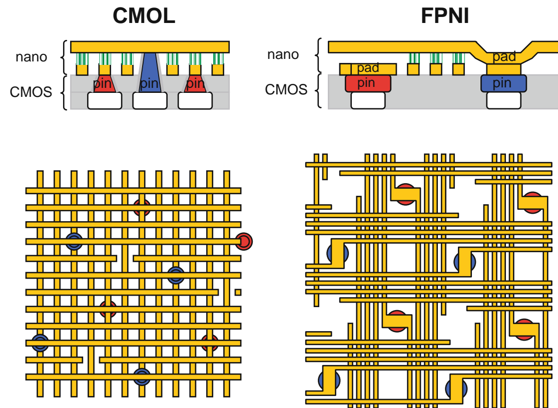


Fig. 4.7 Comparison of CMOL and HP's FPNI circuits (adapted from Ref. [42]) (*See* Color Insert)

degradation is expected to be much less in memories. For example, our preliminary results indicate that density of FPNI-based memory is only about 50% less than that of original CMOL ones [30].

Finally, very recently another promising concept based on CMOL idea was suggested [25]. It is clear from Fig. 4.5 that the most challenging part in the interface is connection to top-layer nanowires. (Actually, the bottom-layer interface can be even further simplified by choosing better pin geometry, e.g., prolonging pin shape along the nanowire direction, without sacrificing the density of the interface.) The suggested modification of CMOL removes this challenging part by placing top-layer interface pins on the other side of crossbar array (Fig. 4.8). This requires stacking of two separately prepared CMOS dies, one with a set of parallel nanowires and device layer on top and another with perpendicular set of parallel nanowires. Clearly, due to the additional CMOS active layer, the performance of such CMOL circuits could be even further improved [24] as compared to the ones based on the original CMOL concept.

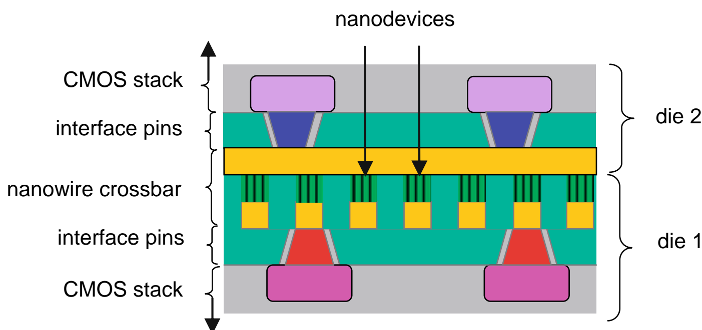


Fig. 4.8 3D CMOL circuits [25]

4.4 CMOL Memories

The most straightforward application of crossbar CMOS/nanodevice hybrids is in memory circuits – see, e.g., theoretical proposals [18, 27, 30, 127] and the first experimental demonstrations [64, 111, 116]. Note that such circuits can be thought of as an extension of more general "crossbar" or "resistive" memory species. In particular, it includes very promising crossbar memories with CMOS-scale wires [56, 65, 128, 129], which have a potential to be the densest among memories based on the conventional photolithography-based technologies. This is why our discussion of these circuits is somewhat relevant for much wider types of memories.

In crossbar memories, nanodevices are used as single-bit memory cells, while the semiconductor transistor subsystem performs all the peripheral (input/output, coding/decoding, line driving, and sense amplification) functions that require relatively smaller number of devices (scaling as $N^{1/2}$, where N is the memory size in bits). If area overhead associated with periphery circuits is negligible then the footprint of the crossbar memories can be as small as $(2F_{\text{nano}})^2$, which might result in the unprecedented density in excess of 1 Tbit/cm^2 at the end of the hybrid technology roadmap (for $F_{\text{nano}} = 3$ nm), i.e., three orders of magnitude higher than that in existing semiconductor memory chips.[5]

The basic operation of crossbar memories can be explained using simplified equivalent circuits shown in Fig. 4.9. In the low-resistive state presenting binary 1, the nanodevice is essentially a diode, so that the application of voltage

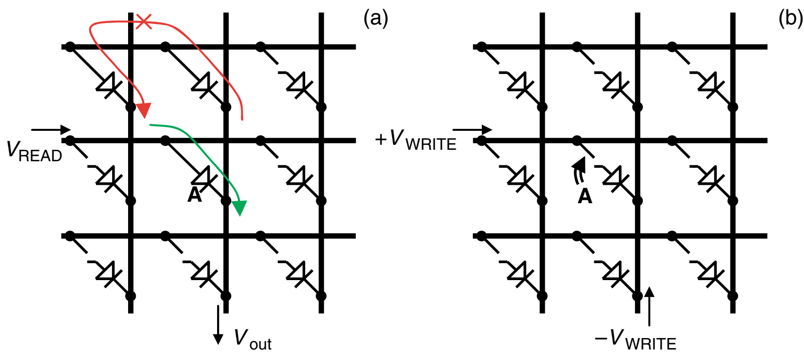


Fig. 4.9 Equivalent circuits of the crossbar memory array showing (**a**) read and (**b**) write operations for one of the cells (marked A). On panel (**a**), *green arrow* shows the useful readout current, while *red arrow* shows the parasitic current to the wrong output wire, which is prevented by the nonlinearity of the $I - V$ curve of device A (if the output voltage is not too high, $V_{\text{out}} < V_t$) (*See* Color Insert)

[5] Here, we do not include in our comparison the data storage systems (such as hard disk drives) which cannot be used for bit-addressable memories because of their very large (millisecond-scale) access time.

$V_t < V_{READ} < V_+$ to one (say, horizontal) nanowire leading to the memory cell gives a substantial current injection into the second wire (Fig. 4.9a). This current pulls up voltage V_{out} which can now be read out by a sense amplifier. The diode property to have low current at voltages above $-V_t$ prevents parasitic currents which might be induced in other state-1 cells by the output voltage – see the red line in Fig. 4.9a. On the other hand, it is easy to show that memory arrays with purely linear (resistive) nanodevices do not scale well and hardly practical [130].

In state 0 (which presents binary zero) the crosspoint current is very small, giving a nominally negligible contribution to output signals at readout. In order to switch the cell into state 1, the two nanowires leading to the device are fed by voltages $\pm V_{WRITE}$ (Fig. 4.9b), with $V_{WRITE} < V_+ < 2V_{WRITE}$. (The left inequality ensures that this operation does not disturb the state of "semiselected" devices contacting just one of the biased nanowires.) The write 0 operation is performed similarly using the reciprocal switching with threshold V_- (Fig. 4.1). It is evident from Fig. 4.9a,b that the read and write operations may be performed simultaneously with all cells of one row.[6]

The main approach for fighting errors in semiconductor memory technology is reconfiguration, i.e., the replacement of memory array lines (rows or columns) containing bad cells by spare lines [131, 132]. The effectiveness of the replacement depends on how good its algorithm is [132, 133]. The Exhaustive Search approach (trying all possible combinations) finds the best repair solution, though it is not practicable because of the exponentially large execution time. A more acceptable choice is the "Repair Most" method that allows a simple hardware implementation and an execution time scaling linearly with the number of bits. In this approach, the number of defects in each line of a memory block (matrix) is counted, and the lines having the largest number of defects are replaced with spare lines.

For a larger fraction of bad bits, better results may be achieved [18, 112, 134] by combining the bad line exclusion with ECC techniques. The simulation results for application of such technique for crossbar hybrid memories [18, 112] have shown that defect tolerance up to $\sim$10% may be achieved using very powerful ECC, e.g., Reed–Solomon and Bose–Chaudhuri–Hocquenghem (BCH) codes [135]. Unfortunately, in those works, the contributions of the circuits implementing these codes to the memory access time (which for some codes may be extremely large) and the total memory area have not been estimated. Also the account of the finite leakage current through nominally closed crosspoints (which was neglected in Ref. [18]) may change the memory scaling rather substantially [117]. What follows is the review of our own approach to terabit-scale defect-tolerant CMOL-based nanoelectronic memories, where we included all relevant overheads in our estimations.

---

[6] Actually, only one of the "write 0" and "write 1" operations can be performed simultaneously with all cells. Because of the opposite polarity of the necessary voltages across nanodevices for these two operations, the complete write may be implemented in two steps, e.g., first writing 0s and then writing 1s.

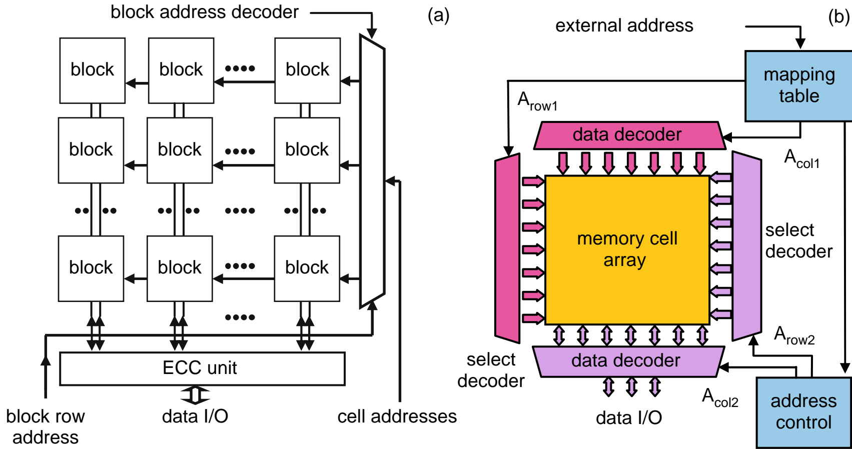


Fig. 4.10 CMOL memory structure: (**a**) global and (**b**) block architectures (*See* Color Insert)

Figure 4.10a shows the assumed general structure of the CMOL memory. Essentially, it is similar to that of the conventional memories, i.e., it is a rectangular array of L crosspoint memory banks ("blocks"), so that during a single operation, a particular row of CMOL blocks is accessed with the help of block address decoders. In contrast, the block architecture (Fig. 4.10b) is specific for the CMOL interface which allows the placement of CMOS "relay" cells under the nanowire crossbar. These cells are controlled by CMOS-level decoders, four per each block (Fig. 4.10b). At each elementary operation, one pair of block decoders (shown in magenta in Fig. 4.10b, as well as in Figs. 4.11 and 4.12a) addresses one vertical and one horizontal CMOS line, and thus selects a certain relay cell at their crosspoint. This cell (Fig. 4.12a) applies the data signal to a "red" interface pin contacting a bottom-layer nanowire. The other pair of decoders (shown in violet in Figs. 4.10b, 4.11, and 4.12a) selects a set of different relay cells which provide similar biasing of the corresponding top-level nanowires through "blue" pins. These nanowires may now address all crosspoint nanodevices (memory cells) of a particular nanowire segment. Thus, the four decoders of the block, working together, can provide every memory cell of the segment with voltages necessary for the read and write operations.

The remaining circuitry shown in Fig. 4.10b, i.e., CMOS-based mapping table and address control circuits, is needed to convert the logical (external) addresses, which are fed to the CMOL blocks, into internal addresses of memory cells inside the block. In particular, the mapping table converts the logical address of the segment (which is the same for all selected blocks) into a pair of block-specific physical addresses, A_{col1} and A_{row1}, and CMOS-implemented decoders activate the corresponding CMOS-level lines.

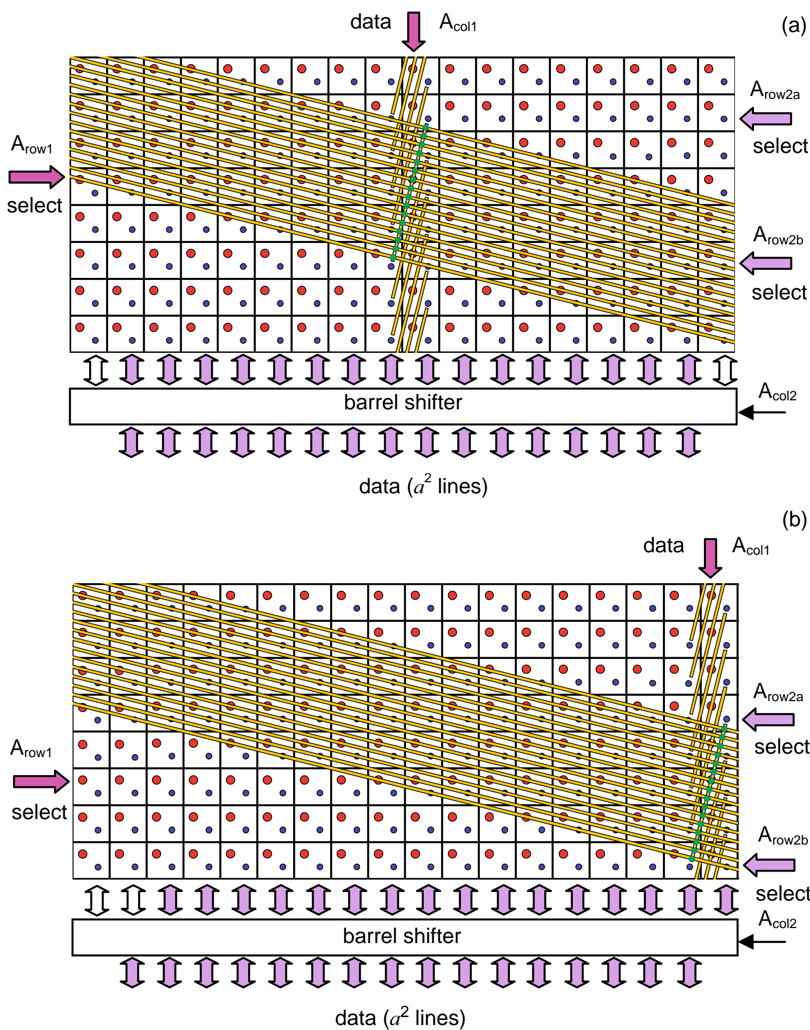


Fig. 4.11 CMOL block architecture: Addressing of an interior column of nanowire segments (for $a = 4$). The figure shows only one (selected) column of the segments, the crosspoint nanodevices connected to one (selected) segment, and continuous top-level nanowires connected to these nanodevices. (In reality, the nanowires of both layers fill all the array plane, with nanodevices at each crosspoint.) The *block arrows* indicate the location of CMOS lines activated at addressing the shown nanodevices (*See* Color Insert)

Figure 4.11 shows the low-level structure of the CMOL memory for a particular (unrealistically small) values of the block size and the main topological parameter of CMOL, $a = 4$. The top-level nanowires (here shown quasi-horizontal) stretch over the whole block, but the low-level

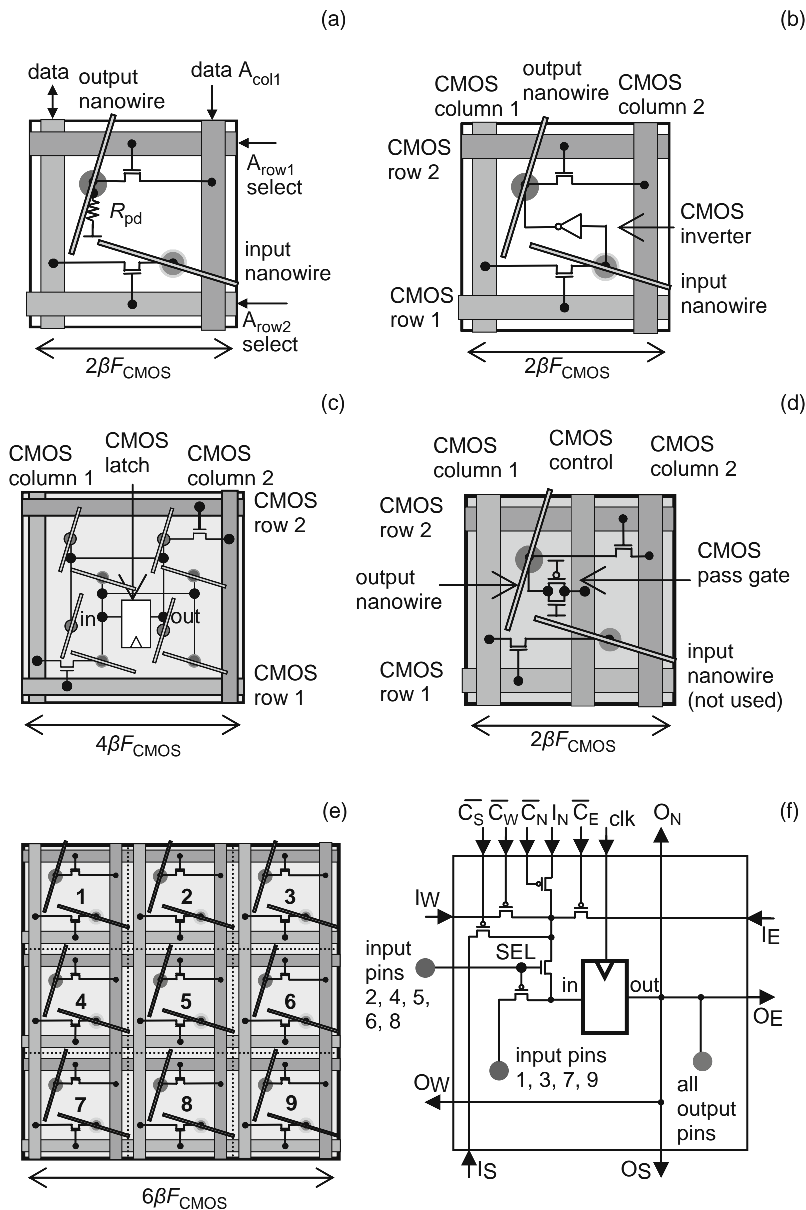


Fig. 4.12 Possible structure of CMOL cells: (**a**) memory relay cell; (**b**) the basic cell: (**c**) the latch cell of CMOL FPGA; (**d**) control cell; (e, f) programmable latch cell of CMOL DSP. Here *red* and *blue points* indicate the corresponding interface pins. For the sake of clarity panels (**a–e**) shows only nanowires which are contacted by interface pins of the given cells. Also for clarity, panel (**e**) shows only the configuration circuitry, while panel (**f**) shows the programmable latch implementation (*See* Color Insert)

(nearly vertical) nanowires are naturally cut into segments of equal length. An elementary analysis of the CMOL geometry (Fig. 4.4) shows that each nanowire segment stretches over a CMOS cells and contacts a^2 (in Fig. 4.11, sixteen) crosspoint nanodevices.

Signals A_{col1} and A_{row1} are applied to CMOS wires, feeding the "red" lines of the corresponding CMOS-implemented relay cells (Fig. 4.12a). By opening all pass transistors of the row, A_{row1} selects a specific "red" pin of column A_{col1}, so that the data A_{col1} are fed only to a specific nanowire segment contacting a^2 crosspoint nanodevices. In parallel, addresses A_{col1} and A_{row1} are sent to the CMOS-based address control circuitry to generate another pair of physical addresses A_{row2} and A_{col2}. Signal A_{row2} opens the "blue"-pin pass transistors in relay cells of a row, and thus connects each of a^2 quasi-horizontal nanowires of the top layer to specific CMOS lines (shown purple), thus enabling a read or write operation.

Our defect tolerance is based on the synergetic approach where memory array reconfiguration is combined with ECC [134].[7] In order to implement this, memory cells are divided into fragments of certain size ("granularity"). Each of these fragments is tested using ECC circuitry, and those of them which may not be ECC corrected are excluded from operation. (For that, the addresses of good fragments are written into the mapping table, see Fig. 4.10b.) If the fraction q of bad bits is large, the large granularity of exclusion is impracticable, due to the exponential growth of the number of necessary redundant resources. On the other hand, fine granularity requires an unacceptably large mapping table. This is why we have used a very flexible approach when the granularity of exclusion is not related to the physical structure of the memory array. This means that the data fragment length, equal to g nanowire segments (i.e., ga^2 memory cells), may be either smaller or larger than the one segment (which has a^2 memory cells).

Requiring that the total yield Y is fixed at a certain level and using detailed performance model [30] we have calculated the total chip area A necessary to achieve a certain useful bit capacity N, and hence the area per useful bit, A/N. The last number, normalized to the CMOS half-pitch area,

$$a \equiv \frac{A}{N(F_{\text{CMOS}})^2}, \tag{4.2}$$

is a very convenient figure of merit that depends only on the ratio $F_{\text{CMOS}}/F_{\text{nano}}$ rather than on the absolute parameters of the fabrication technology.

Figure 4.13 presents typical final results[8] of our optimization procedure, carried out for several values of the total access time. (For our parameters the

[7] We have only considered "stuck-on-open" kind of defects in this work. It is worth mentioning that considered architecture is very efficient for tolerating all other types of defects (e.g., broken or shorted nanowires), except for "stuck-on-close" (permanently shortened) nanodevices.

[8] Though formally the results depend on the total memory size N and yield Y, they are rather insensitive to these parameters in the range of our interest ($N \approx 10^{12}$ bits, $Y \approx 90\%$).

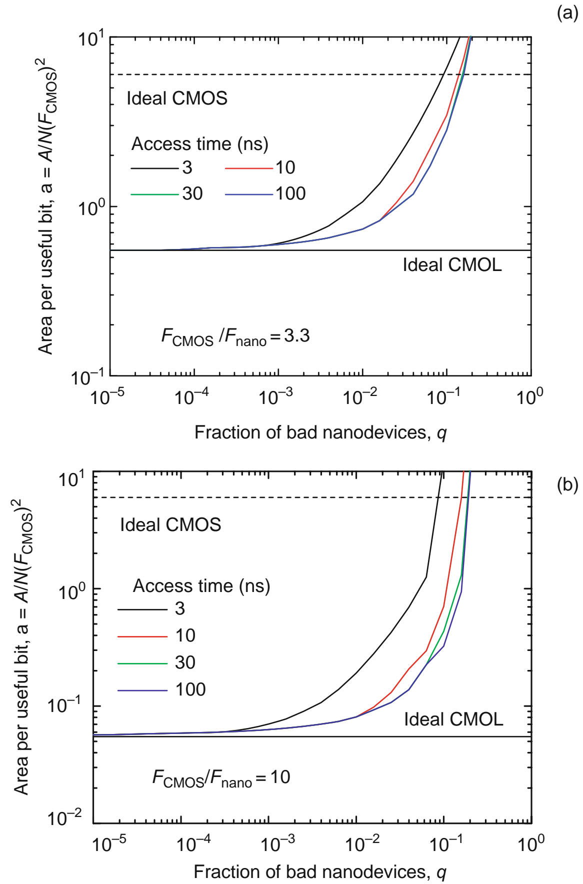


Fig. 4.13 The total chip area per one useful memory cell, as a function of the bad bit fraction q, for several values of the memory access time and two typical values of the $F_{\mathrm{CMOS}}/F_{\mathrm{nano}}$ ratio. The *horizontal lines* indicate the area for "perfect" CMOS and CMOL memories. In the latter case, this line shows our results for negligible q, while for the former case we use the ITRS data [3] for the densest semiconductor (flash) memories (*See* Color Insert)

access time is dominated by the ECC decoding, while intrablock and interblock latencies are negligible.) The cusps on the curves are due to sudden changes of discrete parameters (ga^2, and the number of total and information bits in ECC) for which the largest memory density is achieved. In particular, Fig. 4.13 shows that CMOL memories may become denser than purely CMOS ones at the fraction of bad bit devices as high as ~15 if the latency requirement is not too small (i.e., >10 nm) for both considered cases of pitch ratio. On the other hand, to reach 5× and 10× advantage in density such fraction of bad bits should be below 5% and 2% for $F_{\mathrm{CMOS}}/F_{\mathrm{nano}} = 3.3$ and 10 pitch ratios, correspondingly.

Note, however, that our optimistic results for the memory speed are based on the fundamental physical limitations for the crosspoint nanodevice

---

As Fig. 4.13 shows, the required memory access time τ also has a marginal effect on density, provided τ is not too small.

parameters, in particular, R_{ON}, which was of the order of few kΩs. For the currently implemented programmable diodes, the picture is somewhat different. For example, for the simple and reproducible CuO_x devices [79], scaled down to $F_{nano} = 3$ nm, the effective value of R_{ON} would be ~2 MΩ, resulting in intrablock latency of about 50 ns. This means that our results (Fig. 4.13) would degrade only slightly. On the other hand, for the demonstrated reproducible molecular monolayers [101], typical R_{ON} of a similarly scaled cross-point device would be in the GΩ range, so that the memory latency would be much larger. Nevertheless, a considerable improvement of programmable nanoscale switches during the next decade may be readily anticipated.

4.5 CMOL FPGA Circuits

The practical techniques for high defect tolerance in digital (Boolean) logic are less obvious. In the usual custom logic circuits, the location of a defective gate from outside is hardly possible, while spreading around additional logic gates (e.g., providing von Neumann's majority multiplexing [136]) for error detection and correction becomes very inefficient for fairly low fraction q of defective devices. For example, even the recently improved von Neumann's scheme requires a 10-fold redundancy for q as low as $\sim 10^{-5}$ and a 100-fold redundancy for $q \approx 3 \times 10^{-3}$ [137].

This is why the most significant previously published proposals for the implementation of logic circuits using CMOL-like hybrid structures had been based on reconfigurable regular structures like the field-programmable gate arrays (FPGA). Before this work, two FPGA varieties had been analyzed, one based on look-up tables (LUT) and another one using programmable-logic arrays (PLA).

In the former case, all possible values of an m-bit Boolean function of n binary operands are kept in m memory arrays, of size $2^n \times 1$ each. (For $m = 1$, and some representative applications, the best resource utilization is achieved with n close to 4 [138], while the famous reconfigurable computer Teramac [115] is using LUT blocks with $n = 6$ and $m = 2$.) The main problem with this approach is that the memory arrays of the LUTs based on realistic molecular devices cannot provide address decoding and output signal sensing (recovery). This means that those functions should be implemented in the CMOS subsystem, and the corresponding overhead may be estimated using our results discussed in the previous section. Using the results from Section 4.4, one can show that for the memory array with $2^6 \times 2$ bits, performing the function of a Teramac's LUT block, and for a realistic ratio $F_{CMOS}/F_{nano} = 10$ the area overhead would be above four orders of magnitude (!), and would even loose the density (and hence performance) competition to a purely CMOS circuit performing the same function. On the other hand, increasing the memory array

size to the optimum is not an option because the LUT performance scales (approximately) only as a log of its capacity [138].

The PLA approach is based on the fact that an arbitrary Boolean function can be re-written in the canonical form, i.e., in the two-level logical representation. As a result, it may be implemented as a connection of two crossbar arrays, for example, one performing the AND and another the OR function [33]. The first problem with the application of this approach to the CMOS/nanodevice hybrids is the same as in the case of LUT: the optimum size of the PLA crossbars is finite, and typically small [139], so that the CMOS overhead is extremely large. Moreover, any PLA logic built with diode-like nanodevices faces an additional problem of high power consumption. In contrast to LUT arrays, where it is possible to have current only through one nanodevice at a time, in PLA arrays the fraction of open devices is of the order of one half [36]. Let us estimate the static power dissipated by such an array. The specific capacitance of a wire in an integrated circuit is always of the order of 2×10^{-10} F/m [28]. With $F_{\text{nano}} = 3\,\text{nm}$, this number shows that in order to make the RC time constant of the nanowire below, or of the order of the logic delay in modern CMOS circuits ($\sim 10^{-10}$s), the ON resistance R_{ON} of a molecular device has to be below $\sim 7 \times 10^7$ ohms. For reliable operation of single-electron transistor (and apparently any other active electronic nanodevice) at temperature T, the scale V_{ON} of voltage across it has to be at least $10k_{\text{B}}T$ [1]. For room temperature this gives $V_{\text{ON}} > 0.25$ V, so that static power dissipation per one open device, $P_{\text{ON}} = V_{\text{ON}}^2/R_{\text{ON}}$, is close to 10 nW. With the open device density of $0.5/(2F_{\text{nano}})^2 \approx 10^{12}\ \text{cm}^{-2}$, this creates a power dissipation density of at least $10\,\text{kW/cm}^2$, much higher than the current and prospective technologies allow to manage [3].

As a matter of principle, power consumption may be reduced by using dynamic logic, but this approach requires more complex nanodevices. For example, Refs. [17, 35] describe a dynamic-mode PLA-like structure (with improved functional density via wrapped logic mapping) using several types of molecular-scale devices, most importantly including field-effect transistors which are formed at crosspoints of two nanowires. In such transistor, one (semiconductor) nanowire would serve as a drain/channel/source structure, while the perpendicular nanowire would play the role of the gate. Unfortunately, such circuits would fail because of the same fundamental physical reason that provides the fundamental limitation to the Moore's Law: any semiconductor MOSFET with a few nanometer long channel is irreproducible because of exponential dependence of the threshold voltage on the transistor dimensions [140].[9] Similar problems are likely to prevent hybrid circuits described in Refs. [15, 113] from scaling down beyond 10 nm range, since they are based on nanoscale FETs.

---

[9] In principle, this problem can be alleviated by making the width of nanowires in one dimension comparable with that of lithographically defined wires [35]. However, that also means that such hybrid circuits cannot take full advantage (only in one dimension) of nanodevice nanometer-scale footprints.

Finally, the last significant category of suggested crossbar hybrids includes circuits based on Goto-pair logic [33]. In particular, Refs. [22, 44] describe an architecture where Goto-pair logic is implemented with two-terminal resistive crossbar latches [45]. The main architectural challenge of this approach is due to the fact that nanodevice bistability is employed during Goto-pair operation.[10]

Since the assumed nanodevices have no third state and hence cannot be enabled or disabled, it is unclear how to map a particular circuit on such architectures. (Having a third state would be more challenging since multi-state devices are not very reliable.)

Moreover, the use of bistability in the circuit operation is rather impractical due to the relation between the retention time and the switching speed in the crossbar latches. In order to be useful for most electronics applications, the latches should be switched very fast (in a few picoseconds in order to compete with advanced MOSFETs), but retain their internal state for the time necessary to complete the calculation (ideally, for a few years, though several hours may be acceptable in some cases). This means that the change of the applied voltage by the factor of 2 (the difference between the fully selected and semiselected crosspoints of a crossbar) should change the switching rate by at least 16 orders of magnitude. However, even the most favorable physical process we are aware of (the quantum-mechanical tunneling through high-quality dielectric layers like the thermally grown SiO_2) may only produce, at these conditions, the rate changes below 10 orders of magnitude, even if uncomfortably high voltages of the order of 12 V are used [141].

Let us now discuss an alternative approach to Boolean logic circuits based on CMOL concept [26, 28, 29] that is closed to the so-called cell-based FPGA [142]. We have studied two varieties of CMOL FPGA fabrics [26, 28, 29]. The architecture of the simplest variety, one-cell fabric [28], is very convenient for elaborating the concept and basic properties of CMOL FPGA, though it cannot be used for sequential circuit design. On the other hand, a two-cell fabric [26, 29], which is a generalization of the single-cell structure, can be used for mapping arbitrary circuits, and all simulation results in this section will be given for such a variety of CMOL FPGAs.

Figure 4.14 shows a fragment of one-cell CMOL FPGA fabric. Essentially, it is a uniform structure which is built by replicating "basic" cells with an area $A = (2\beta F_{\mathrm{CMOS}})^2$. In this case, the angle α is given by the generic formula for CMOL, i.e., $\tan\alpha \equiv 1/a$ (Eq. 4.1), where a is an integer defining the range of cell interaction (Fig. 4.14).[11] For fixed fabrication technology parameters F_{CMOS},

---

[10] As a reminder, in all discussed crossbar circuits above, as well as in our approach for Boolean logic described in this section and Section 4.6, the state of nanodevices remains unchanged during circuit operation.

[11] Note that even though the nanowire crossbar in Ref. [28] was rotated by the additional 45° angle, which was convenient for manual mapping, it does not affect the performance results.

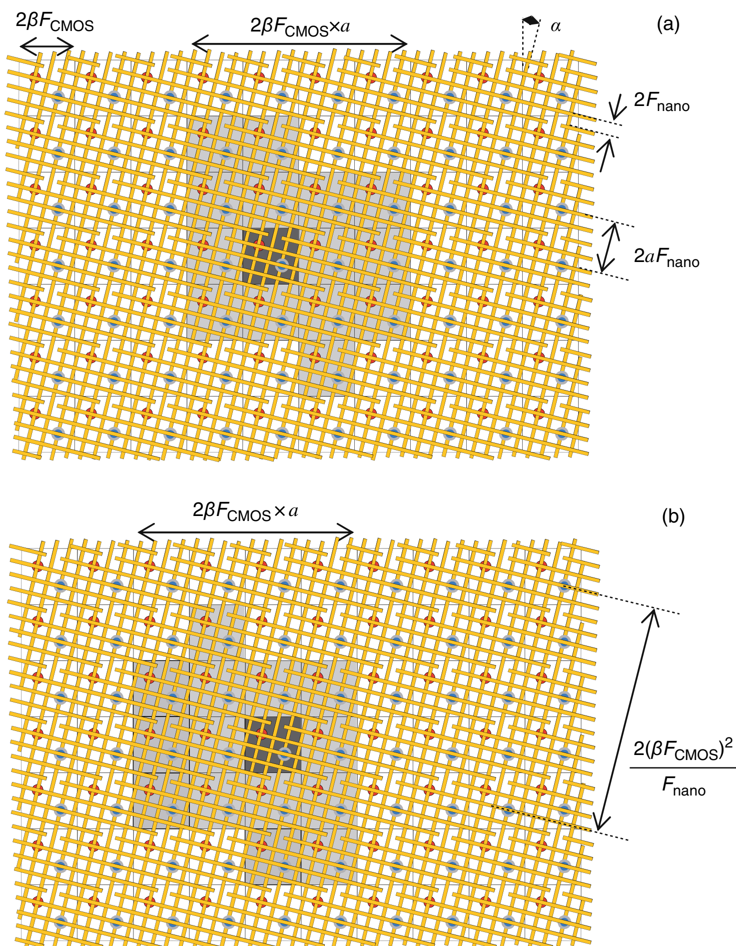


Fig. 4.14 The fragment of one-cell CMOL FPGA fabric for the particular case $a = 4$. In panel (**a**), output pins of $M = a^2 - 2 = 14$ cells (which form the so-called input cell connectivity domain) painted *light gray* may be connected to the input pin of a specific cell (shown *dark gray*) via a pin-nanowire-nanodevice-nanowire-pin links. Similarly, panel (**b**) shows cells (painted *light gray*) whose inputs may be connected directly to the output pin of a specific cell (called output connectivity domain) (*See* Color Insert)

F_{nano}, and $\beta_{\min}$, the lower bound on a is given by inequality:

$$a^2 > (\beta_{\min} F_{\text{CMOS}}/F_{\text{nano}})^2 - 1. \tag{4.3}$$

Each basic cell (Fig. 4.12b) consists of an inverter and two pass transistors that serve two pins (one of each type) serving as the cell input and output, respectively. During the configuration stage, all inverters are disabled by an

appropriate choice of global voltages V_{DD} and V_{gnd} (Fig. 4.12b), and testing and setting of all nanodevices is carried out absolutely similarly to memory read/write operation described in the previous section.

In contrast to CMOL memories, nanowires in upper layer are also fabricated with small breaks repeated with period $L = 2(\beta F_{CMOS})^2/F_{nano}$. With this arrangement, each nanowire segment is connected to exactly one interface pin.[12] As a result, each input or output of a basic cell can be connected through a pin-nanowire-nanodevice-nanowire-pin link to each of

$$M = a^2 - 2 \tag{4.4}$$

other cells located within a square-shaped "cell connectivity domain" around the initial cell – see Fig. 4.14. (For infinitesimal gaps, M would equal $a^2 - 1$, but for a more feasible gap width of the order of $2F_{nano}$, the connectivity domain is by one cell smaller.) Note that in reality both input and output cell connectivity domains would be much larger than those shown in Fig. 4.14 for practical values of $a > 10$ and have the same roughly square shape (with some protrusions of the cells on the perimeter of the domain). This fact simplifies the design automation for CMOL FPGA circuits [26].

When the configuration stage has been completed, the pass transistors are used as pull-down resistors, while the nanodevices set into ON (low-resistive) state are used as pull-up resistors. Together with CMOS inverters, these components may be used to form the basic "wired-NOR" gates (Fig. 4.15). For example, if only the two nanodevices shown in a Fig. 4.15b are in the ON state, while all other latching switches connected to the input nanowire of cell H are in the OFF (high resistance) state, then cell H calculates NOR function of signals A and B. Clearly, the gates with high fan-in (Fig. 4.15c) and fan-out or broadcast (Fig. 4.15d) may be readily formed as well as by turning ON the corresponding latching switches. Having these primitives is sufficient to implement any Boolean function, as well as to perform routing, provided the hardware resources are sufficient.

A genuine optimization of CMOL FPGA circuit architectures would require a completely new set of CAD tools, whose development is a challenging task. At this preliminary stage, our choice was instead to get as much leverage as possible from the existing ideas and algorithms used for mapping and architecture exploration of semiconductor logic, in particular, from the design automation algorithms for island-type CMOS FPGAs [143].

---

[12] The best performance is achieved if the pin contacts the wire fragment in its middle, and our analysis has been carried out with this assumption. Since lower layer nanowire segments are cut by upper layer pins, a connection exactly in a center is easily achievable, i.e., by locating upper level pins correspondingly. For upper layer pins, a similar trick can be done, if upper layer nanowire breaks are provided by features of the same lithographic mask that defines interface pin positions. Also note that a modest misalignment of the pin and the breaks (by $\sim F_{CMOS}$) reduces the circuit performance only by a small factor of the order of $1/\beta \ll 1$.

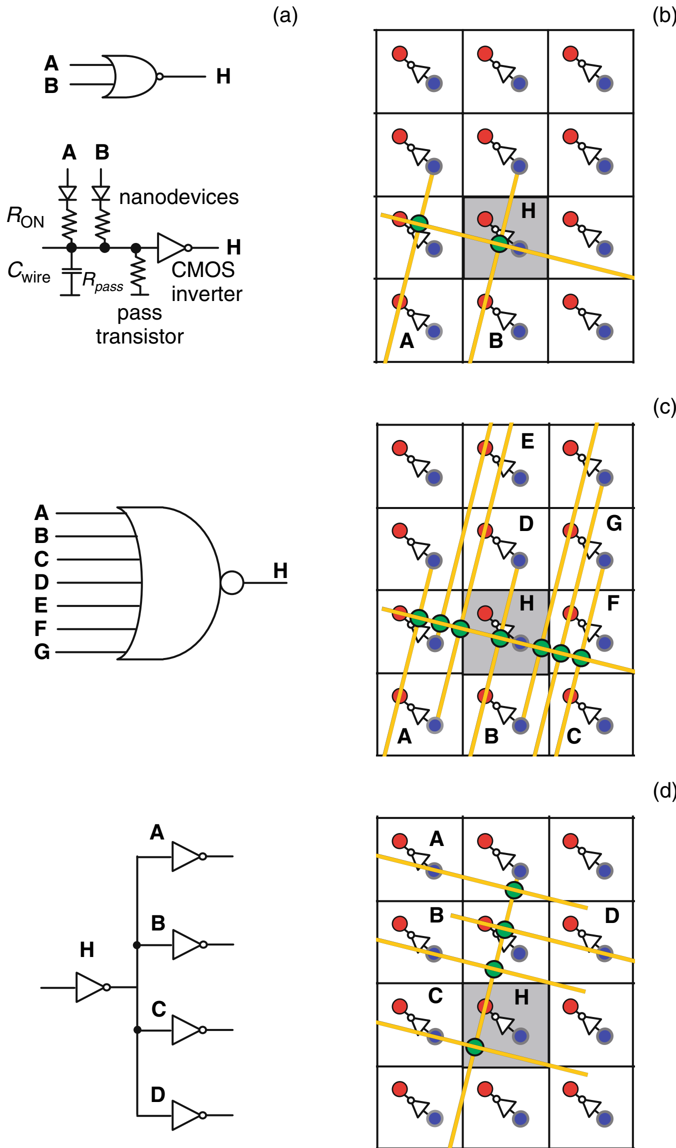


Fig. 4.15 Logic and routing primitives in CMOL FPGA circuits: (**a**) equivalent circuit of fan-in-two NOR gate, (**b**) its physical implementation in CMOL, (**c**) the example of 7-input NOR gate, and (**d**) the example of fan-out of signal to four cells. Note that only several (shown) nanodevices on the input nanowires in panels (**b**), (**c**), and output nanowire in panel (**d**) of cell H are set to the ON state, while others (not shown) are set to the OFF state. Also, for the sake of clarity, panels (**b**)–(**d**) show only the nanowires used for the gate and the broadcast (*See* Color Insert)

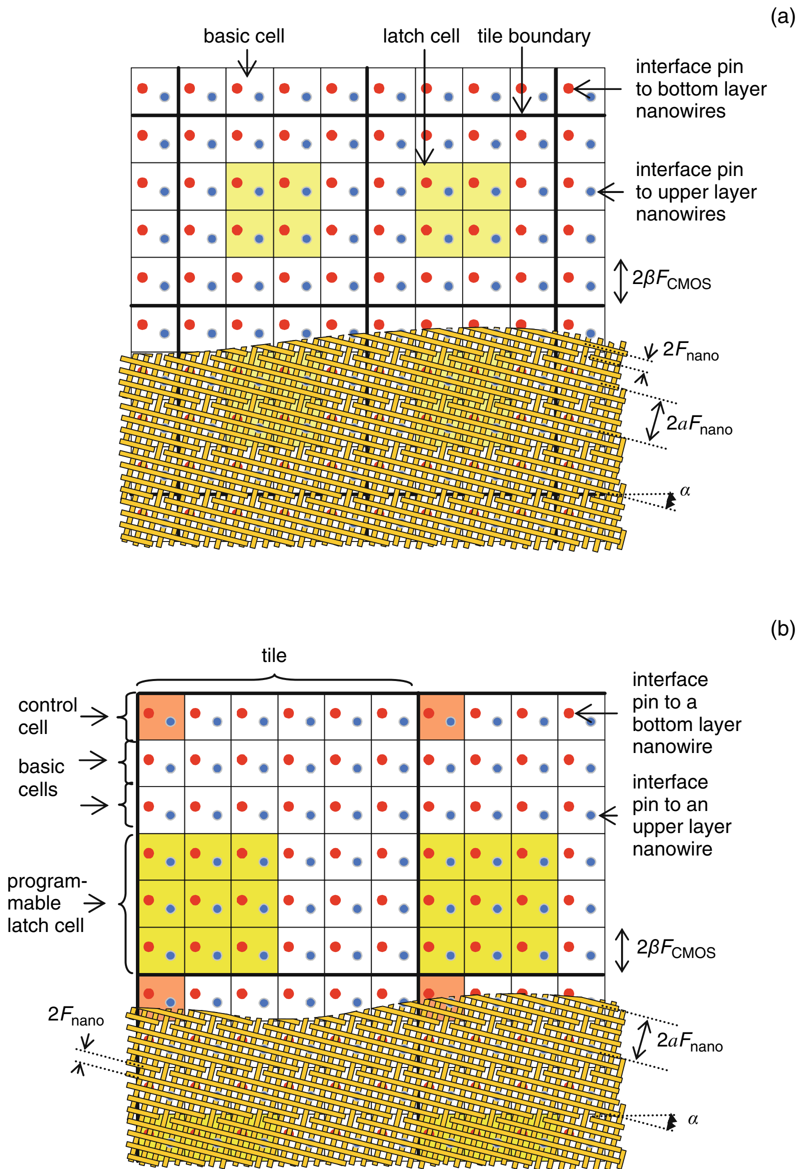


Fig. 4.16 A fragment of (**a**) two-cell CMOL FPGA fabric and (**b**) three-cell CMOL DSP fabric for the particular case $a = 4$ (*See* Color Insert)

In order to use such design automation algorithms, we have restricted our design to a specific, simple two-cell-species CMOL fabric. The fabric is a uniform mesh of square-shaped "tiles" (Fig. 4.16a). Each tile consists of a shell of T basic cells (Fig. 4.12c) surrounding a single "latch" cell (Fig. 4.12d).

The latter cell is just a level-sensitive latch implemented in the CMOS subsystem, connected to eight interface pins, plus two pass transistors used for circuit configuration. Note that all four pins of each (either input or output) group are always connected, so that the nanowires they contact always carry the same signal. This means that at configuration, groups of four nanodevices sitting on these wires may be turned on or off only together. A simple analysis shows that this does not impose any restrictions on the CMOL FPGA fabric functionality.

The convenience of the proposed two-cell CMOL FPGA structure is that, from the design point of view, the CMOL tile can be treated in the same way as that of the island-type CMOS FPGA. To demonstrate that let us first introduce a very useful concept of "tile connectivity domain" which makes routing of CMOL FPGA circuits similar to CMOS FPGA ones. Similar to the cell connectivity domain, the tile connectivity domain of a given tile is defined as such fabric fragment that any cell within it can be connected to any cell of the initial tile directly, i.e., via one pin-nanowire-nanodevice-nanowire-pin link (Fig. 4.17). Just as for cell connectivity domains all tile connectivity domains

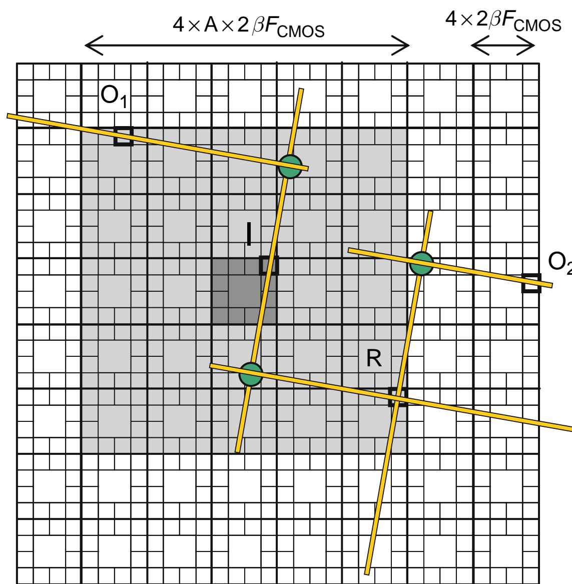


Fig. 4.17 Tile connectivity domain: Any cell of the central tile (shown *dark gray*) can be connected with any cell in the tile connectivity domain (shown *light gray*) via one pin-nanowire-nanodevice-nanowire-pin link (e.g., cells I and O_1). Cells outside of each other's tile connectivity domain (e.g., I and O_2) can be connected with additional routing inverters (e.g., R). Note that nanowire width and nanodevice size are boosted for clarity. For example, for the considered CMOL parameters, 1600 crosspoint nanodevices may fit in one basic cell area (*See* Color Insert)

are similar and have square shape. (Note that we assume that input and output tile connectivity domains are the same.) The linear size $\mathbb{A}$ of the tile connectivity domain for the assumed tile size $T = 16$ can be found as

$$\mathbb{A} = 2\lfloor a/8 \rfloor - 1. \tag{4.5}$$

For instance, Fig. 4.17 shows a tile connectivity domain for the case $A = 5$. (In more realistic cases $a = \beta F_{\mathrm{CMOS}}/F_{\mathrm{nano}} \approx 40$, i.e., $\mathbb{A} \approx 9$.)

The main idea of the proposed design flow for CMOL FPGAs (Fig. 4.18) is to reserve some number of basic cells ($T - K$) inside each tile for routing purposes, while use the rest of the cells (K) for logic during the placement step. The placer tries to put gates into such locations (with maximum one latch and K NOR gates per tile) so that their interconnect is local or, equivalently, is within tile connectivity domain of each other. At the global routing step, idle cells inside each tile are used to interconnect global connections. If there is a congestion after the global routing step, i.e., the number of requested basic cells during routing $count_{\max}$ is larger than the actual number of idle cells $T - K - \bar{\Delta}$ (here $\bar{\Delta}$ is parameter which allows to trade off the number of iterations with the mapping quality), then we decrease K and repeat the flow again until there is no congestion.

We have applied our methods to analyze possible CMOL FPGA implementation of the Toronto 20 benchmark circuit set [144]. Using the completely custom design automation flow [26, 29], we have first mapped the circuits on the two-cell CMOL FPGA fabric. Then, assuming a plausible power supply

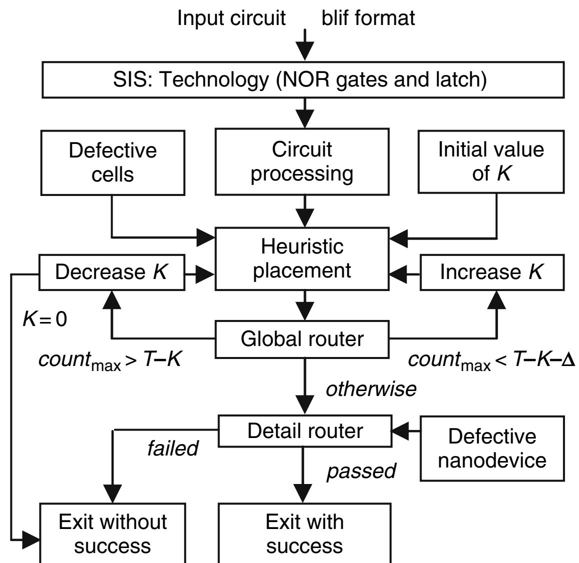


Fig. 4.18 CMOL FPGA design flow used in this work

voltage $V_{DD} = 0.3\,V$ [29], we find the smallest acceptable nanodevice resistances, and consequently the lowest delay of CMOL gates, such that the power consumption density does not exceed the ITRS-specified value of $200\,W/cm^2$ and the voltage swing on the input of CMOS inverters is sufficiently larger than the corresponding shot and thermal noise of nanodevices [28, 29]. From such optimization, the crosspoint resistance in the ON state is about $280\,k\Omega$, while delay τ_0 of a NOR-1 gate is about 80 ps.

Table 4.1 summarizes the performance results for the benchmark circuits mapped on CMOL FPGA without any defects. Note that in contrast to earlier nanoelectronics work, the results for different circuits are obtained for the CMOL FPGA fabric with exactly the same operating conditions and physical structure for all the circuits, thus enabling a fair comparison with CMOS FPGA. For this comparison, the same benchmark circuits have been synthesized into cluster-based island-type logic block architecture [143] and scaled (using very optimistic assumptions) to get CMOS FPGA performance for similar CMOS technology node [28]. Also,

Table 4.1 Performance results for Toronto 20 benchmark set mapped on two-cell CMOL fabric with no defects

	CMOS FPGA ($F_{CMOS} = 45\,nm$)		CMOL FPGA ($F_{CMOS} = 45\,nm$, $F_{nano} = 4.5\,nm$, max fan-in = 7)		Comparison	
Circuit	Area (μm^2)	Delay (ns)	Area (μm^2)	Delay (ns)	A_{CMOS} / A_{CMOL}	$A_{nanoPLA}$ / A_{CMOL}
alu4	137,700	5.1	1,004	4.0	137	0.28
apex2	166,050	6.0	914	4.6	182	3.09
apex4	414,619	5.5	672	3.6	617	0.58
bigkey	193,388	3.1	829	2.7	233	1.82
clma	623,194	13.1	9,308	10.2	67	1.74
des	148,331	4.2	1,097	4.5	135	3.21
diffeq	100,238	6.0	1194	10.4	84	2.27
dsip	148,331	3.2	829	3.4	179	1.63
elliptic	213,638	8.6	4,581	12.7	47	1.63
ex1010	391,331	9.0	3,486	5.7	112	0.28
ex5p	100,238	5.1	829	4.3	121	0.19
frisc	230,850	11.3	4,199	17.6	55	2.64
misex3	124,538	5.3	1,004	3.6	124	0.56
pdc	369,056	9.6	4,979	6.8	74	0.15
s298	166,050	10.7	829	8.1	200	1.33
s38417	462,713	7.3	9,308	7.2	50	1.24
s38584	438,413	4.8	9,872	8.8	44	–
seq	151,369	5.4	1,296	4.0	117	1.15
spla	326,025	7.3	2,994	5.8	109	0.12
tseng	78,469	6.3	1,194	11.5	66	2.48

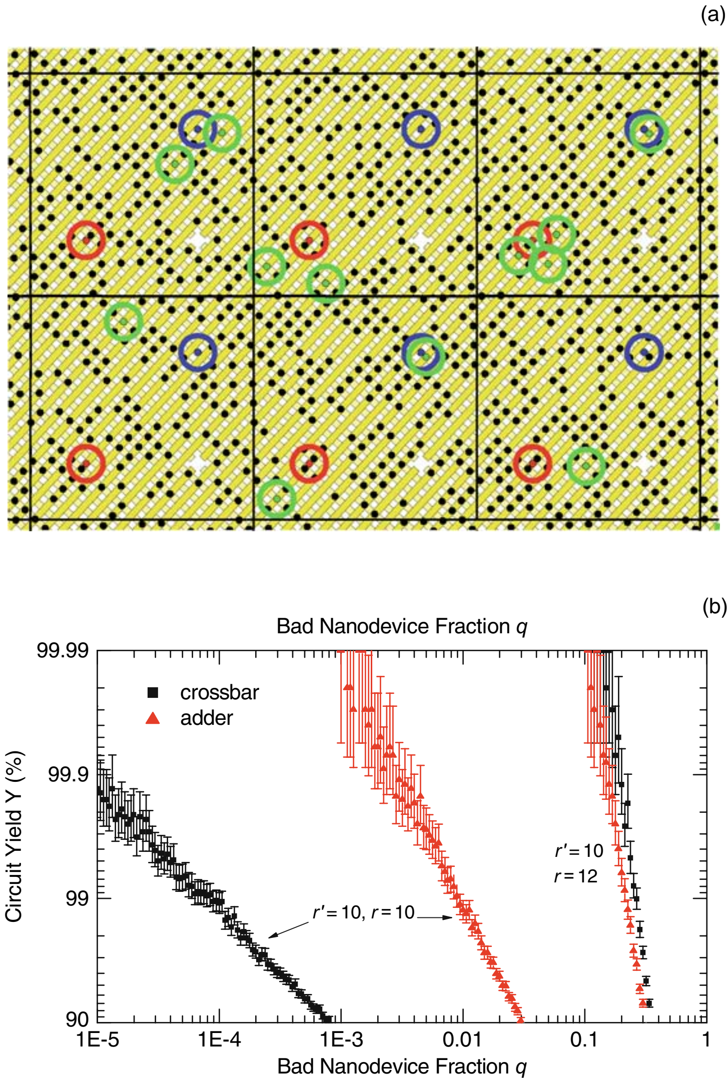


Fig. 4.19 (**a**) A small fragment of the 32-bit Kogge-Stone adder mapped on one-cell CMOL fabric after the reconfiguration as around 50% stuck-on-open nanodevices. Bad nanodevices are shown *black*, good used devices *green*, unused devices are not shown, for clarity. *Colored circles* are only a help for the eye, showing the location of interface pins (*red* and *blue* points) and used nanodevices. *Thin vertical* and *horizontal lines* show CMOS cell borders. (**b**) The final (post-reconfiguration) defect tolerance of 32-bit Kogge-Stone adder and the 64-bit full crossbar for several values of $F_{\mathrm{CMOS}}/F_{\mathrm{nano}}$. For more details – see Ref. [28] (*See* Color Insert)

for a comparison we show in Table 1 the simulation results for NanoPLA concept presented in Ref. [118].

Finally, our simulations have shown that the CMOL FPGA is very resilient to various types of defects [26, 28]. For example, Fig. 4.19b shows that at realistic parameter $a = 40$ circuits may have a fabrication yield above 99% with up to 20% of stuck-on-open nanodevices [28]. Such high defect tolerance should not be surprising because only a small percent of nanodevices (about 0.1% of the total on average) is utilized. A huge redundancy can be efficiently exploited by the detail routing algorithm, which allows to pick completely deferent set of nanodevices by moving positions of the gates [28]. Indeed, in some cases, successful reconfiguration around as many as 50% of bad nanodevices is possible (Fig. 4.19a). Also, our simulations have shown that the CMOL FPGA is very resilient to defective CMOS cells [26]. In particular, the average swelling of the circuit area is rather limited: only about 20% and 80% for 10% and 30%, respectively, uniformly distributed defective cells (see, e.g., Fig. 4.20). This means that faulty interface pins, nanowires, and/or CMOS circuitry can be very effectively tolerated. On the other hand, the tolerance to stuck-on-close crosspoint defects is rather low (i.e., equivalent to about 0.02% of defective nanodevices for 30% defective cells) so that some other defect tolerance mechanism should be used to reduce the effects of such faults.

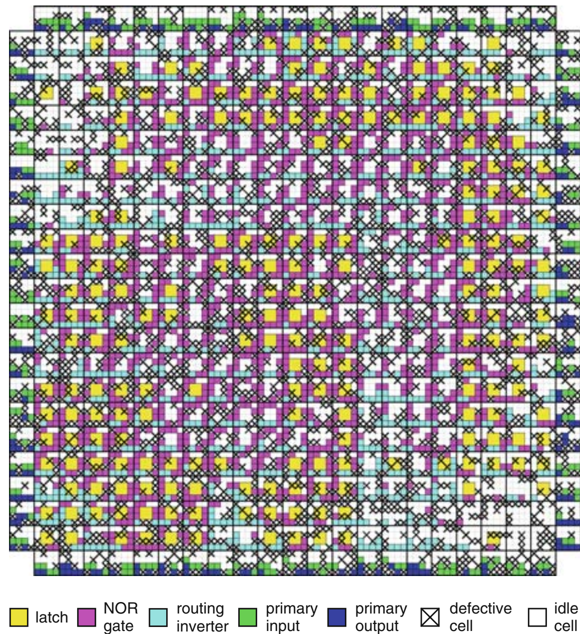


Fig. 4.20 Example of mapping on two-cell CMOL fabric with a presence of defective cells: dsip.blif circuit of the Toronto 20 set, mapped on the (21 + 2)×(21 + 2) tile array with 30% defective cells. Here the additional layer of tiles at the array periphery is used exclusively for I/O functions. The cells from these peripheral tiles are functionally similar to input and output pads and cannot be configured to NOR gates (*See* Color Insert)

4.6 CMOL DSP Circuits

In general, any type of Boolean logic circuits can be mapped on CMOL FPGAs. With several modifications, CMOL FPGA can be turned into a circuit architecture which is especially efficient for a very important class of applications – low-level image processing tasks. (Such low-level tasks are performed frequently in spatial filtering, edge detection, feature extraction, etc., and typically present a bottleneck in image processing systems.) More specifically, let us discuss possible performance of such circuits on a simple but representative example of 2D image convolution:

$$T_{x,y} = \sum_{i=1}^{F} \sum_{j=1}^{F} S_{x+i,\ y+j} \varphi_{i,\ j}, \tag{4.6}$$

where S and T are input and output images, correspondingly, with $N \times N$ pixels each, and φ is a $F \times F$ pixel filter function. Though sometimes special rules for calculating the edge pixels of image T are used [4], we will consider a simplified version of the algorithm where the linear size of output image is smaller by $(F - 1)$ pixels (Fig. 4.21), so that all output signals T are calculated according to Eq. (4.6). Such simplification should not affect the performance results for more general case since, typically, $F \ll N$. For example, the baseline parameters used for estimates in this work are $F = 32$, $N = 1024$, with the similar accuracy ($n_S = n_T = n_\varphi = n = 12$ bits) of the input, output, and filter data.

Figure 4.22 shows the top-level architecture of the proposed CMOL-based DSP. Here we assume the most challenging I/O option when the data are fed to

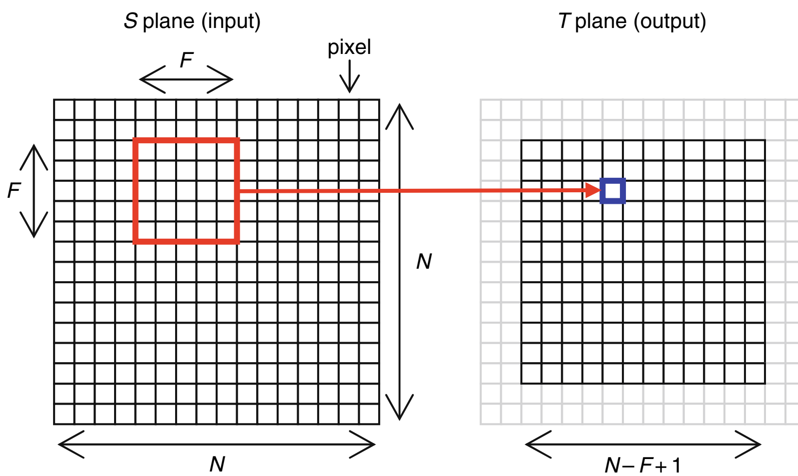


Fig. 4.21 Scheme of the 2D image convolution for particular (impracticably small) sizes of the initial image ($N = 16$) and filter window ($F = 5$)

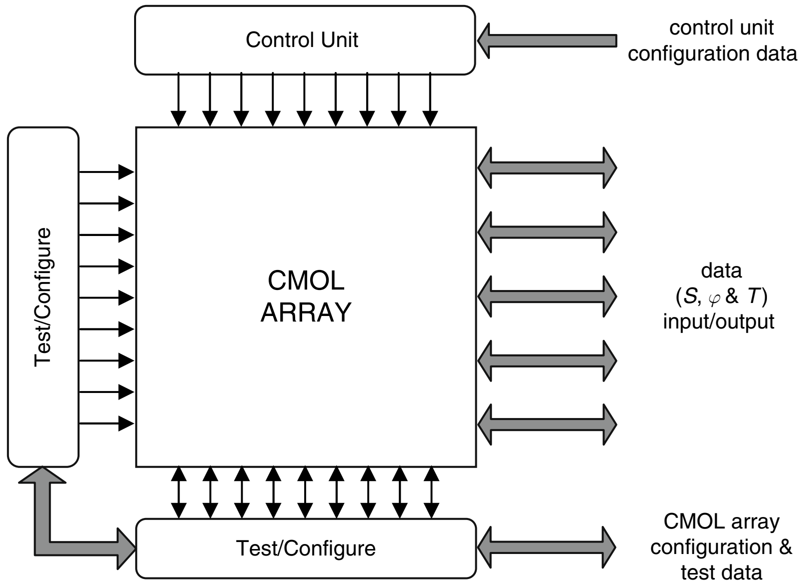


Fig. 4.22 The top-level structure of the CMOL DSP

and picked up from the array from the periphery, e.g., the right side on Fig. 4.22. (If an area-distributed 2D I/O is available, the circuit performance would only be better.) The key part of the architecture, the CMOL array, is similar for each pixel. The pixel area is organized into a uniform mesh of square-shaped "tiles" (Fig. 4.16). The number of tiles per pixel depends on the data word length *n*; in our case it is close to 12×12. Each tile consists of 26 basic cells, one control cell of the similar size, and one programmable latch cell of a larger area (Fig. 4.16b). Just like in circuits discussed above, the CMOS circuitry of each cell type is different (see Fig. 4.12b, d, e, f), but the interface and nanowire levels are the same for all cell types: similarly located pins of each elementary cell contact nanowire fragments of the same length. The CMOS-implemented configuration circuitry, comprised of a pair of signal lines (shown in blue and magenta in Fig. 4.12b, d, e) and a pass transistor per pin, is also similar for each cell.

There are two new types of cells. The control cell (Fig. 4.12d) connects its output nanowire to the CMOS control unit memory (outside the CMOL array) via the designated CMOS control lines [8]. The programmable latch cell (Fig. 4.12e, f) is designed to provide not only temporary data storage but also a fast (CMOS-wire) interconnect between each tile and its four nearest neighbors – see inputs I_S, I_N, I_W, I_E and outputs O_S, O_N, O_W, O_E in Fig. 4.12f. The latter feature may be used both for window operations and for a fast transfer of data in and out the CMOL array. To implement these functions, each latch cell has a CMOS latch with a programmable input. More specifically, depending on the value of signal SEL, which may arrive from any of five input nanowires

(Fig. 4.12e), the input of CMOS latch can be connected to either any of other four input nanowires or one of the neighboring latch cells. The particular choice of the neighbor is determined by the signals arriving via CMOS-implemented select lines C_S, C_N, C_W, C_E, which are common to all programmable latch cells. CMOS layout estimates have shown that control cells can be fit into the $64(F_{CMOS})^2$ squares each, thus giving $\beta_{min} = 4$. The programmable latch readily fits into an area nine times larger.

It is obvious from Eq. (4.6) that the convolution can be effectively parallelized. For example, the convolution process may be broken into F^2 sequential steps, each corresponding to a specific pair of indices i, j, the same for all pixels. At each step, every pixel with coordinates x, y of the CMOL array is supplied with one component $S_{x+i,y+j}$ of the input signal matrix, and all pixels are supplied, in parallel, with the same component $\varphi_{i,j}$ of the window function. During the step, the pixel circuitry calculates the product $S_{x+i,y+j}\varphi_{i,j}$ and adds it to the partial sum of $T_{x,y}$, which is kept in that pixel all the time. These add-and-multiply operations are done in all pixels in parallel (Fig. 4.23), so that the whole convolution (of one input frame) is accomplished in F^2 steps.

One of the advantages of the considered parallelization algorithm is that all pixels may have similar structure (Fig. 4.24). In this schematics, the two most complex parts are the 12-bit multiplier (for the partial product generation and reduction) and the 32-bit adder. The adder is used both for the last step in the multiplication and for the summation of the products in Eq. (4.6). Such dual use of the adder requires additional multiplexers (whose CMOL-DSP implementation is described in details in Ref. [8]) at the input of the adder (i.e., cA and cB in Fig. 4.24) and latch M for keeping intermediate values.

We have found [8] that the best performance for the convolution task with the considered parameters can be achieved with a multiplier featuring straightforward partial product generation and the Wallace-tree-like reduction scheme [145]. The summation is implemented with a parallel 32-bit Kogge-Stone adder, designed in our previous work [28]. The remaining pixel hardware (not shown in Fig. 4.24) includes bypass circuitry – see Ref. [8] for more discussion. Figure 4.25 shows the typical pixel mapping. Though the utilization is very high, i.e., about 80% of the whole area of the pixel, there is still some space to add more functions to the pixel (e.g., a more sophisticated rounding scheme) if necessary, without increasing its area.

We have estimated the performance of the 2D image convolution mapped on CMOL DSP chip for a particular choice of parameters, $F_{nano} = 4.5$ nm and $F_{CMOS} = 45$ nm, which might be typical for the initial stage of the CMOL technology development [146]. Using the cell area estimates made in previous sections, the size of one pixel is about $25\times25\ \mu m^2$, while the size of full CMOL array with $N = 1024$ pixels is $\sim 25\times25\ mm^2$.

Our latency calculations also followed those of previous works on digital CMOL circuits [28, 29]. Because of slightly higher nanodevice utilization as

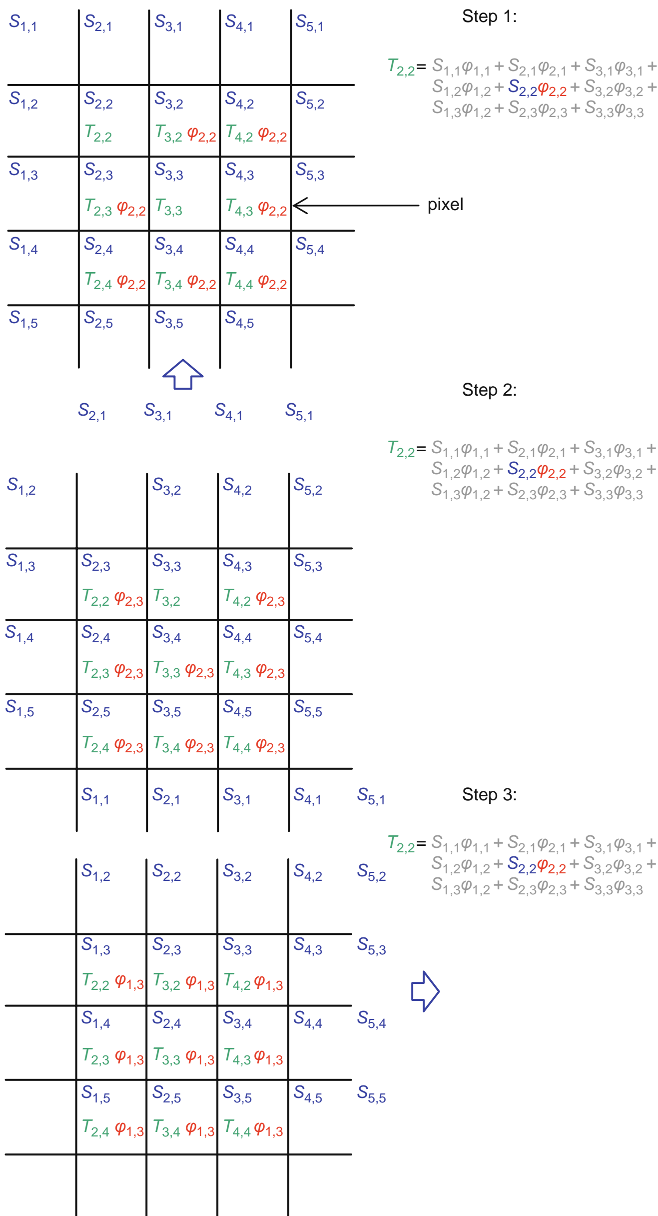


Fig. 4.23 Three sequential time steps of the convolution in the *left top* corner of the CMOL DSP array for $F = 3$. Colored terms in the formulas below each panel show the calculated partial sums in the pixel 2,2. For the (uncharacteristically small) filter size, it takes just $F^2 = 9$ steps to complete the processing of one frame (*See* Color Insert)

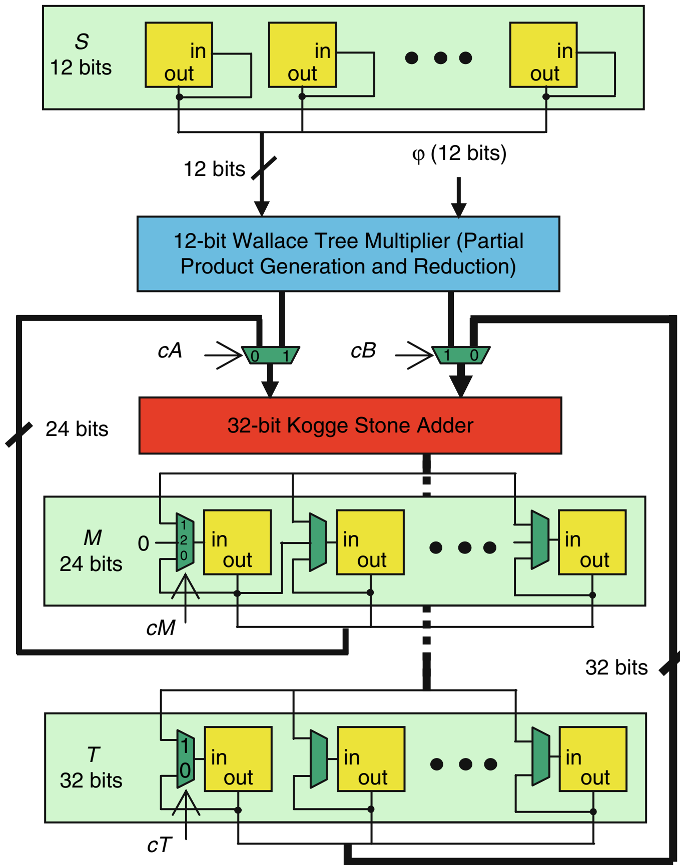


Fig. 4.24 Pixel schematics for the 2D convolution with considered parameters. Note that this picture only shows connections which are implemented with nanowires. In reality there are also connections (between neighbor programmable latches) implemented with CMOS-scale lines

compared to circuits in the previous section in CMOL DSP the delay τ_0 of a NOR-1 gate is about 100 ps. Including the delay of shifting data in and out of the array the full delay for our parameters may be estimated as 25 μs. Also, our simulation has shown that a similar defect tolerance to that of previous section is plausible for CMOL DSP circuits [8].

Even assuming a very optimistic (linear) delay scaling and possible increase in the number of cores (from 8 to 32), the corresponding latency of a hypothetical 45 nm 6.4 GHz Cell processor would be about 3.5 ms [8]. This number is at least 100 times larger than that of proposed CMOL DSP. This is not surprising since the CMOL DSP has a much higher peak performance. For example, it can theoretically perform 250×10^{12} 32-bit additions per second or about 100×10^{12} 12-bit multiply–add operations per second, i.e., above two

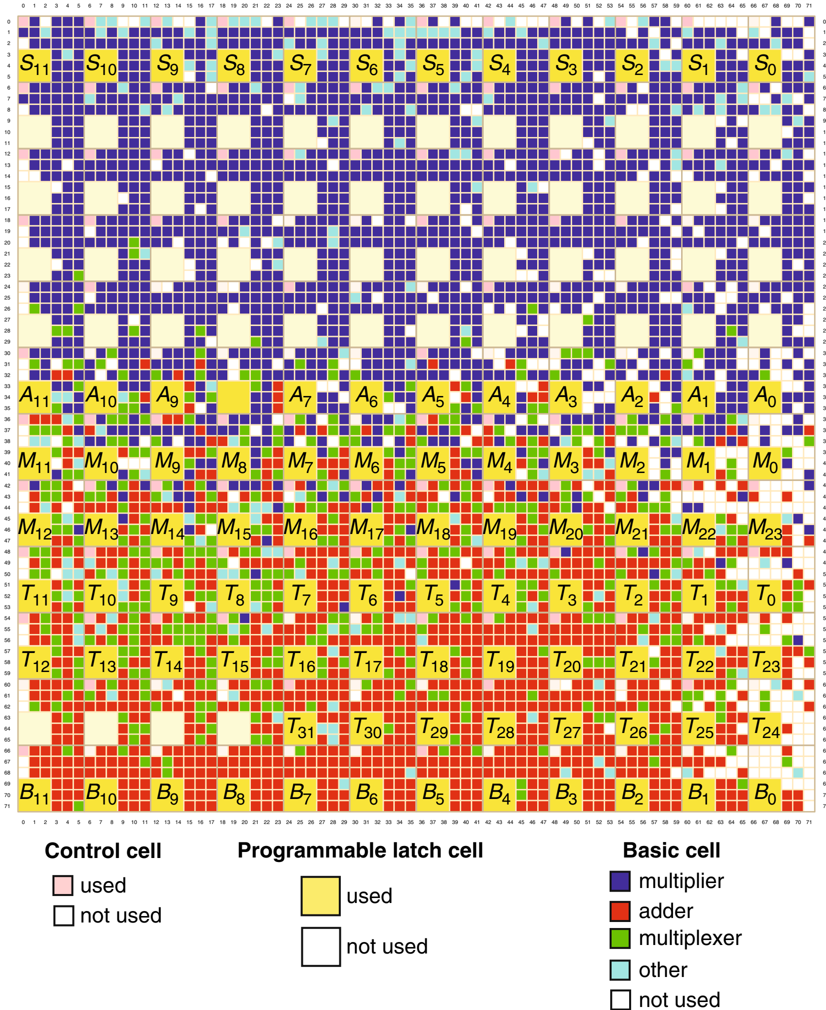


Fig. 4.25 The mapping of the pixel on CMOL DSP (for $F_{\text{CMOS}}/F_{\text{nano}} = 10$) after its successful reconfiguration of the circuit around as many as 40% of bad nanodevices with random locations. Programmable latches A and B are used for bypass circuitry during the data up and down shift operations (*See* Color Insert)

orders of magnitude higher than Cell. Moreover, the theoretical data bandwidth of a CMOL DSP could be as high as 10 Tbit/s, which should be enough for even very demanding applications.[13]

---

[13] It is worth noting that a major advantage of the Cell-type processors for the low-level image processing tasks is a very fast (nanosecond-scale) time necessary for changing the running task (e.g., the filter size). CMOL DSP can almost certainly have a sub-100 μs time of switching from one task to another.

4.7 Conclusions

Simulation results presented in this work clearly show that CMOL-based digital circuits may continue the performance scaling of microelectronics well beyond the limits of currently dominating CMOS technology. We believe that these prospects more than justify large-scale research and development efforts in the synthesis of functional molecular devices, their chemically directed self-assembly, nanowire patterning, and CMOL circuit architectures. In particular, the range of urgent hardware development tasks includes the following [146]

- The design, fabrication, and characterization of programmable diodes
- Scaling of reproducible crosspoint nanodevices below 10 nm (which may require the transfer from the metal oxide-based programmable diodes to single-electron-based SAM junctions [1, 32])
- Experimental demonstration of an area-distributed CMOL interface, which may radically change the industrial perception of the hybrid CMOS/nano-device circuits

Acknowledgments The author is especially grateful to K. K. Likharev who equally contributed to the results presented in this chapter. Also, useful discussions of various aspects of digital CMOL circuits with J. Barhen, V. Beiu, R. Brayton, S. Chatterjee, S. Das, A. DeHon, D. Hammerstrom, A. Korkin, P. Kuekes, J. Lukens, A. Mayr, A. Mishchenko, N. Quitoriano, G. Snider, M. Stan, D. Stewart, N. H. Di Spigna, R. S. Williams, T. Zhang, and N. Zhitenev are gratefully acknowledged. The work has been supported in part by AFOSR, DTO, and NSF.

References

1. K. K. Likharev, Electronics below 10 nm, in *Nano and Giga Challenges in Microelectronics*, pages 27–68, Elsevier, Amsterdam, 2003.
2. D. J. Frank et al., Proc. IEEE **89**, 259 (2001).
3. available online at http://public.itrs.net/.
4. R. Lewis, *Practical Digital Image Processing*, Ellis Horwood, New York, 1990.
5. P. H. Swain and S. M. Davis, *Remote Sensing. The Quantitative Approach*, McGraw-Hill, New York, 1978.
6. S. M. Chai et al., Appl. Optics **39**, 835 (2000).
7. D. C. Pham et al., IEEE J. Solid-State Circ. **41**, 179 (2006).
8. D. B. Strukov and K. K. Likharev, IEEE Tran. Nanotechnol. (2007), accepted for publication.
9. D. B. Strukov and K. K. Likharev, IEEE Tran. Nanotechnol. **6**, 696 (2007).
10. C. A. Stafford, D. M. Cardamone, and S. Mazumdar, Nanotechnology **18**, 424014 (2007).
11. W. Wang, T. Lee, and M. Reed, Intrinsic electronic conduction mechanisms in self-assembled monolayers, in *Introducing Molecular Electronics*, edited by G. Cuniberti, G. Fagas, and K. Richter, pages 275–300, Springer, Berlin, 2005.
12. D. Porath, DNA-based devices, in *Introducing Molecular Electronics*, edited by G. Cuniberti, G. Fagas, and K. Richter, pages 411–446, Springer, Berlin, 2005.

13. S. C. Goldstein and M. Budiu, NanoFabrics: Spatial computing using molecular electronics, in *Proc. of ISCA'01*, pages 178–189, GÖteborg, Sweden, 2001.
14. M. Masoumi, F. Raissi, M. Ahmadian, and P. Keshavarzi, Nanotechnology **17**, 89 (2006).
15. T. Wang, Z. Qi, and C. A. Moritz, Opportunities and challenges in application-tuned circuits and architectures based on nanodevicess, in *Proc. of CCF'04*, pages 503–511, Italy, 2004.
16. M. M. Ziegler and M. R. Stan, IEEE Trans. Nanotechnol. **2**, 217 (2003).
17. A. DeHon, IEEE Trans. Nanotechnol. **2**, 23 (2003).
18. A. DeHon, S. C. Goldstein, P. J. Kuekes, and P. Lincoln, IEEE Trans. Nanotechnol. **4**, 215 (2005).
19. M. M. Ziegler et al., Molecular Electronics III **1006**, 312 (2003).
20. T. Hogg and G. Snider, JETTA **23**, 117 (2007).
21. G. Snider, P. Kuekes, T. Hogg, and R. S. Williams, Appl. Phys. A-Mater. Sci. Process. **80**, 1183 (2005).
22. G. S. Snider and P. J. Kuekes, IEEE Trans. Nanotechnol. **5**, 129 (2006).
23. T. Hogg and G. S. Snider, IEEE Trans. Nanotechnol. **5**, 97 (2006).
24. C. Dong, W. Wang, and S. Haruehanroengra, Micro & Nano Lett. **1**, 74 (2007).
25. D. Tu, M. Liu, W. Wang, and S. Haruehanroengra, Micro & Nano Lett. **2**, 40 (2007).
26. D. B. Strukov and K. K. Likharev, CMOL FPGA circuits, in *Proc. of CDES'06*, pages 213–219, Las Vegas, NE, 2006.
27. D. B. Strukov and K. K. Likharev, Nanotechnology **16**, 137 (2005).
28. D. B. Strukov and K. K. Likharev, Nanotechnology **16**, 888 (2005).
29. D. B. Strukov and K. K. Likharev, A reconfigurable architecture for hybrid CMOS/Nanodevice circuits, in *Proc. of FPGA*'06, pages 131–140, New York, ACM Press, 2006.
30. D. B. Strukov and K. K. Likharev, J. Nanosci. Nanotechnol. **7**, 151 (2007).
31. A. DeHon and K. Likharev, Hybrid CMOS/nanoelectronic digital circuits: Devices, architectures, and design automation, in *Proc. of ICCAD'05*, pages 375–382, 2005.
32. K. K. Likharev and D. B. Strukov, CMOL: Devices, circuits, and architectures, in *Introducing Molecular Electronics*, edited by G. Cuniberti, G. Fagas, and K. Richter, pages 447–478, Springer, Berlin, 2005.
33. M. R. Stan, P. D. Franzon, S. C. Goldstein, J. C. Lach, and M. M. Ziegler, Proc. IEEE **91**, 1940 (2003).
34. P. J. Kuekes, G. S. Snider, and R. S. Williams, Sci. Am. **293**, 72 (2005).
35. A. DeHon, ACM J. Emerg. Technol. Comput. Syst. **1**, 109 (2005).
36. S. Das, G. Rose, M. M. Ziegler, C. A. Picconatto, and J. E. Ellenbogen, Architecture and simulations for nanoprocessor systems integrated on the molecular scale, in *Introducing Molecular Electronics*, edited by G. Cuniberti, G. Fagas, and K. Richter, pages 479–515, Springer, Berlin, 2005.
37. K. K. Likharev, Hybrid semiconductor/nanoelectronic circuits, in *Proc. of NanoTech'07*, pages 552–555, Cambridge, MA, 2007.
38. K. K. Likharev, A. Mayr, I. Muckra, and Ö. Türel, Ann. NY Acad. Sci. **1006**, 146 (2003).
39. O. Türel, J. H. Lee, X. L. Ma, and K. K. Likharev, Int. J. Circ. Theory App. **32**, 277 (2004).
40. J. H. Lee and K. K. Likharev, Int. J. Circ. Theory App. **35**, 239 (2007).
41. G. S. Snider, Nanotechnology **18**, 365202 (2007).
42. G. S. Snider and R. S. Williams, Nanotechnology **18**, 035204 (2007).
43. A. A. Gayasen, N. Vijaykrishnan, and M. J. Irwin, Exploring technology alternatives for nano-scale FPGA interconnects, in *Proc. of DAC'05*, pages 921–926, 2005.
44. G. Snider, Appl. Phys. A-Mater. Sci. Process. **80**, 1165 (2005).
45. P. J. Kuekes, D. R. Stewart, and R. S. Williams, J. Appl. Phys. **97**, 034301 (2005).
46. D. R. Stewart et al., Nano Lett. **4**, 133 (2004).
47. G. Dearnaley, A. M. Stoneham, and D. V. Morgan, Reports on Progress in Phys. **33**, 1129 (1970).

48. G. Cuniberti, G. Fagas, and K. Richter, editors, *Introducing Molecular Electronics*, Springer, Berlin, 2005.
49. J. C. Scott and L. D. Bozano, Adv. Mater. **19**, 1452 (2007).
50. A. M. Bratkovsky, Current rectification, switching, polarons, and defects in molecular electronics devices, in *Polarons in Advanced Materials*, edited by A. S. Alexandrov, Canopus/Springer, Bristol, England, 2007.
51. H. Pagnia and N. Sotnik, Phys. Status Solidi A-Appl. Res. **108**, 11 (1988).
52. L. P. Ma, S. Pyo, J. Ouyang, Q. F. Xu, and Y. Yang, Appl. Phys. Lett. **82**, 1419 (2003).
53. L. D. Bozano, B. W. Kean, V. R. Deline, J. R. Salem, and J. C. Scott, Appl. Phys. Lett. **84**, 607 (2004).
54. J. Y. Ouyang, C. W. Chu, C. R. Szmanda, L. P. Ma, and Y. Yang, Nature Mater. **3**, 918 (2004).
55. F. Verbakel, S. C. J. Meskers, and R. A. J. Janssen, Appl. Phys. Lett. **89**, 102103 (2006).
56. R. Sezi et al., IEDM Tech. Digest, 10.2.1 (2003).
57. L. P. Ma, Q. F. Xu, and Y. Yang, Appl. Phys. Lett. **84**, 4908 (2004).
58. Y. S. Lai, C. H. Tu, D. L. kWong, and J. S. Chen, Appl. Phys. Lett. **87**, 122101 (2005).
59. Q. X. Lai, Z. H. Zhu, Y. Chen, S. Patil, and F. Wudl, Appl. Phys. Lett. **88**, 133515 (2006).
60. J. H. A. Smits, S. C. J. Meskers, R. A. J. Janssen, A. W. Marsman, and D. M. de Leeuw, Adv. Mater. **17**, 1169 (2005).
61. C. P. Collier et al., Science **285**, 391 (1999).
62. J. H. Krieger, S. V. Trubin, S. B. Vaschenko, and N. F. Yudanov, Synthetic Metals **122**, 199 (2001).
63. C. Li et al., App. Phys. Lett. **82**, 645 (2003).
64. W. Wu et al., App. Phys. A-Mater. Sci. Process. **80**, 1173 (2005).
65. Y.-C. Chen et al., IEDM Tech. Digest, 37.4.1 (2003).
66. M. Kund et al., IEDM Tech. Digest, 31.5 (2005).
67. Z. Wang et al., IEEE Elec. Dev. Lett. **28**, 14 (2007).
68. H. B. Chung, K. Shin, and J. M. Lee, J. Vac. Sci. Technol. A **25**, 48 (2007).
69. M. H. R. Lankhorst, B. W. S. M. M. Ketelaars, and R. A. M. Wolters, Nature Mater. **4**, 347 (2005).
70. M. N. Kozicki, M. Park, and M. Mitkova, IEEE Trans. Nanotechnol. **4**, 331 (2005).
71. J. Hu, A. J. Snell, J. Hajto, M. J. Rose, and W. Edmiston, Thin Solid Films **396**, 240 (2001).
72. J. M. Shannon, S. P. Lau, A. D. Annis, and B. J. Sealy, Solid-State Electron. **42**, 91 (1998).
73. C. A. Richter, D. R. Stewart, D. A. A. Ohlberg, and R. S. Williams, Appl. Phys. A-Mater. Sci. Process. **80**, 1355 (2005).
74. J. R. Jameson et al., Appl. Phys. Lett. **91**, 112101 (2007).
75. B. J. Choi et al., J. Appl. Phys. **98**, 033715 (2005).
76. D. S. Jeong, H. Schroeder, and R. Waser, Electrochem. Solid State Lett. **10**, G51 (2007).
77. K. M. Kim, B. J. Choi, Y. C. Shin, S. Choi, and C. S. Hwang, Appl. Phys. Lett. **91**, 012907 (2007).
78. H. Sim et al., Microelectron. Eng. **80**, 260 (2005).
79. A. Chen et al., IEDM Tech. Digest, 31.3 (2005).
80. D. C. Kim et al., Appl. Phys. Lett. **88**, 202102 (2006).
81. H. Shima et al., Appl. Phys. Lett. **91**, 012901 (2007).
82. S. Seo et al., Appl. Phys. Lett. **85**, 5655 (2004).
83. H. Shima, F. Takano, Y. Tamai, H. Akinaga, and I. H. Inoue, Jap. J. Appl. Phys. 2 **46**, L57 (2007).
84. G. Stefanovich, A. Pergament, and D. Stefanovich, J. Phys.-Cond. Matter **12**, 8837 (2000).
85. B. G. Chae, H. T. Kim, D. H. Youn, and K. Y. Kang, Phys. B-Cond. Matt. **369**, 76 (2005).
86. A. Asamitsu, Y. Tomioka, H. Kuwahara, and Y. Tokura, Nature **388**, 50 (1997).
87. K. Szot, W. Speier, G. Bihlmayer, and R. Waser, Nature Mater. **5**, 312 (2006).

88. D. S. Kim, Y. H. Kim, C. E. Lee, and Y. T. Kim, Phys. Rev. B **74**, 174430 (2006).
89. A. Sawa, T. Fujii, M. Kawasaki, and Y. Tokura, Appl. Phys. Lett. **88**, 232112 (2006).
90. S. Karg, G. I. Meijer, D. Widmer, and J. G. Bednorz, Appl. Phys. Lett. **89**, 072106 (2006).
91. R. Fors, S. I. Khartsev, and A. M. Grishin, Phys. Rev. B **71**, (2005).
92. M. Hamaguchi, K. Aoyama, S. Asanuma, Y. Uesu, and T. Katsufuji, Appl. Phys. Lett. **88**, 142508 (2006).
93. J. R. Contreras et al., Appl. Phys. Lett. **83**, 4595 (2003).
94. P. Levy et al., Phys. Rev. B **65**, 140401 (2002).
95. C. N. Lau, D. R. Stewart, R. S. Williams, and M. Bockrath, Nano Lett. **4**, 569 (2004).
96. D. S. Jeong, H. Schroeder, and R. Waser, Appl. Phys. Lett. **89**, 082909 (2006).
97. J. J. Blackstock et al., APS Meeting Abstracts, 10004 (2006).
98. J. Blanc and D. L. Staebler, Phys. Rev. B **4**, 3548 (1971).
99. S. FÖlling, Ö. Türel, and K. K. Likharev, Single-electron latching switches as nanoscale synapses, in *Proc. of IJCNN'01*, pages 216–221, Mount Royal, New York, Int. Neural Network Soc., 2001.
100. N. B. Zhitenev, H. Meng, and Z. Bao, Phys. Rev. Lett. **88**, 226801 (2002).
101. H. B. Akkerman, P. W. M. Blom, D. M. de Leeuw, and B. de Boer, Nature **441**, 69 (2006).
102. J. Goldberger, A. I. Hochbaum, R. Fan, and P. D. Yang, Nano Lett. **6**, 973 (2006).
103. T. Bryllert, L. E. Wernersson, T. Lowgren, and L. Samuelson, Nanotechnology **17**, S227 (2006).
104. C. M. S. Torreset al., vMater. Sci. Eng. C-Biomimetic Supramol. Syst. **23**, 23 (2003).
105. L. J. Guo, J. Phys. D-Appl. Phys. **37**, R123 (2004).
106. D. J. Wagner and A. H. Jayatissa, Nanoimprint lithography: Review of aspects and applications, in *Nanofabrication: Technologies, Devices, and Applications II*, edited by W. Y. Lai, L. E. Ocola, and S. Pau, volume 6002, page 60020R, SPIE, 2005.
107. I. W. Hamley, Angewandte Chemie-International Edition **42**, 1692 (2003).
108. S. R. J. Brueck, There are no fundamental limits to optical lithography, in *International Trends in Applied Optics*, pages 85–109, SPIE Press, Bellingham, WA, 2002.
109. H. H. Solak et al., Microelectron. Eng. **67–8**, 56 (2003).
110. G.-Y. Jung et al., Nano Lett. **2**, 351 (2006).
111. Y. Chen et al., Nanotechnology **14**, 462 (2003).
112. F. Sun and T. Zhang, IEEE Trans. Nanotechnol. **6**, 341 (2007).
113. G. Snider, P. Kuekes, and R. S. Williams, Nanotechnology **15**, 881 (2004).
114. A. DeHon and H. Naeimi, IEEE Des. Test Comput. **22**, 306 (2005).
115. J. R. Heath, P. J. Kuekes, G. S. Snider, and R. S. Williams, Science **280**, 1716 (1998).
116. J. E. Green et al., Nature **445**, 414 (2007).
117. C. J. Amsinck, N. H. Di Spigna, D. P. Nackashi, and P. D. Franzon, Nanotechnology **16**, 2251 (2005).
118. A. DeHon, Design of programmable interconnect for sublithographic programmable logic arrays, in *Proc. of FPGA*'05, pages 127–137, Monterey, CA, 2005.
119. S. C. Goldstein and D. Rosewater, Digital logic using molecular electronics, in *Proc. of ISSCC'02*, page 12.5, San Francisco, CA, 2002.
120. A. DeHon, P. Lincoln, and J. E. Savage, IEEE Trans. Nanotechnol. **2**, 165 (2003).
121. P. J. Kuekes, W. Robinett, G. Seroussi, and R. S. Williams, App. Phys. A-Mater. Sci. Process. **80**, 1161 (2005).
122. P. J. Kuekes et al., Nanotechnology **17**, 1052 (2006).
123. B. Gojman, E. Rachlin, and J. E. Savage, ACM J. Emerg. Technol. Comput. Syst. **1**, 73 (2005).
124. G. F. Cerofolini, Appl. Phys. A-Mater. Sci. Process. **86**, 31 (2007).
125. K. K. Likharev, Interface **14**, 43 (2005).
126. N. H. Di Spigna, D. P. Nackashi, C. J. Amsinck, S. R. Sonkusale, and P. Franzon, IEEE Trans. Nanotechnol. **5**, 356 (2006).

127. R. J. Luyken and F. Hofmann, Nanotechnology **14**, 273 (2003).
128. N. E. Gilbert and M. N. Kozicki, IEEE J. Solid-State Circ. **42**, 1383 (2007).
129. I. G. Baek et al., IEDM Tech. Digest, 31.4 (2005).
130. P. P. Sotiriais, IEEE Tran. Inf. Theory **52**, 3019 (2006).
131. B. Prince, *Semiconductor Memories: A Handbook of Design, Manufacture, and Application*, Wiley, Chichester, 2nd edition, 1991.
132. K. Chakraborty and P. Mazumder, *Fault-Tolerance and Reliability Techniques for High-Density Random-Access Memories*, Prentice Hall, Upper Saddle River, NJ, 2002.
133. C. T. Huang, C. F. Wu, J. F. Li, and C. W. Wu, IEEE Trans. Reliab. **52**, 386 (2003).
134. C. H. Stapper and H. S. Lee, IEEE Trans. Comput. **41**, 1078 (1992).
135. R. E. Blahut, *Algebraic Codes for Data Transmission*, Cambridge University Press, Cambridge, 2003.
136. J. von Neumann, Probabilistic logics and the synthesis of reliable organisms from unreliable components, in *Automata Studies*, edited by G. Cuniberti, G. Fagas, and K. Richter, pages 329–378, Princeton Univeristy Press, Princeton, NJ, 1956.
137. S. Roy and V. Beiu, IEEE Trans. Nanotechnol. **4**, 441 (2005).
138. E. Ahmed and J. Rose, IEEE Trans. VLSI **12**, 288 (2004).
139. J. Kouloheris and A. E. Gamal, PLA-based FPA versus cell granularity, in *Proc. of CICS'92*, pages 4.3.1–4, Boston, MA, 1992.
140. V. A. Sverdlov, T. J. Walls, and K. K. Likharev, IEEE Trans. Electron Dev. **50**, 1926 (2003).
141. W. D. Brown and J. Brewer, *Nonvolatile semiconductor memory technology: a comprehensive guide to understanding and to using NVSM devices*, IEEE Press series on microelectronic systems., IEEE Press, New York, 1998.
142. J. M. Rabaey, A. P. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits: A Design Perspective*, Pearson Education, Upper Saddle River, NJ, 2nd edition, 2003.
143. V. Betz, J. Rose, and A. Marquardt, *Architecture and CAD for deep-submicron FPGAs*, Kluwer Int. Series in Eng. and Comp. Science 497, Kluwer Academic, Boston, London, 1999.
144. *FPGA place-and-route challenge*, 1999, Available online at http://www.eecg.toronto.edu/~vaughn/challenge/ challenge.html/.
145. M. J. Flynn and S. F. Oberman, *Advanced Computer Arithmetic Design*, Wiley, New York, 2001.
146. K. K. Likharev and D. B. Strukov, Prospects for the development of digital CMOL circuits, in *Proc. of NanoArch'07*, pages 109–116, San Jose, CA, 2007.

Chapter 5
Fundamentals of Spintronics in Metal and Semiconductor Systems

Roland K. Kawakami, Kathleen McCreary, and Yan Li

5.1 Introduction

Spintronics is a new paradigm for electronics which utilizes the electron's spin in addition to its charge for device functionality [1, 2]. The primary areas for applications or potential applications are information storage, computing, and quantum information. In terms of materials, the study of spin in solids now includes metallic multilayers [1], inorganic semiconductors [2, 3], transition metal oxides [4, 5], organic semiconductors [6, 7, 8, 9, 10], and carbon nanostructures [11, 12, 13, 14, 15, 16, 17, 18, 20, 21, 22]. The diversity of materials studied for spintronics is a testament to the advances in synthesis, measurement, and interface control that lie at the heart of nano-electronics. In terms of technology, the discoveries of giant magnetoresistance (GMR) [23], tunneling magnetoresistance (TMR) [24], and spin torque [25] in metallic multilayers have led to significant advances in high-density hard drives and non-volatile random access memory. Advances in semiconductor spintronics [3] including the observation of long spin coherence times and the demonstration of spin manipulation point toward potential applications in advanced computation such as reconfigurable logic and quantum information processing.

This tutorial is intended for graduate students just starting in the field of spintronics and for experts in other fields. Our goal is to cover some of the key ideas in spintronics and to present the material in an intuitive and pedagogical manner. We provide "back of the envelope" calculations whenever possible. Due to the quantum mechanical nature of spin, some of the calculations require a working knowledge of quantum mechanics, so a short appendix is provided to review some of the quantum mechanical properties of spin.

R.K. Kawakami
Department of Physics and Astronomy, Center for Nanoscale Science and Engineering, University of California, Riverside, CA 92521
e-mail: roland.kawakami@ucr.edu

A. Korkin, F. Rosei (eds.), *Nanoelectronics and Photonics*,
DOI: 10.1007/978-0-387-76499-3_5, © Springer Science+Business Media, LLC 2008

After a brief introduction to some important properties of spin, the remainder of the chapter is organized on the following lines of research:

- Metallic magnetic multilayers
- Semiconductor spintronics
- Lateral spin transport devices

The selection of topics represents areas in which the authors have had first-hand experience and areas closely related to those.

While the ultimate technological impacts of spintronics cannot be predicted, it is a subject worth learning due to its current technological successes and the potential for revolutionary technologies in the future. It is hoped that some readers will be motivated to learn more about spintronics and perhaps contribute to this rapidly evolving enterprise.

5.1.1 Why Spin?

We begin with the question, "What is special about electron spin?" From a scientific and technological point of view, there are four important points. First is the connection between spin and magnetism, which is useful for information storage. Second is an intrinsic connection between spin and quantum mechanics, which may be useful for quantum information. Third is the short range of spin-dependent exchange interactions, which implies that the role of spin will continue to grow as the size of nanostructures continues to shrink. Fourth are the issues of speed and power dissipation, which are becoming increasingly important for electronics at the nanoscale.

First, spin is connected to ferromagnetic materials because the spontaneous magnetization breaks time-reversal symmetry, which allows the electronic states within the material to become spin dependent. This contrasts with non-magnetic materials where time-reversal symmetry forces the electronic states to come in pairs with the same energy but opposite spin (Kramer's degeneracy), thus leading to a density of states that must be independent of spin. Figure 5.1 shows a schematic diagram of the density of states for a ferromagnetic metal and a non-magnetic metal. In the ferromagnetic metal, the density of states is different for the two spin states. It is conventional to refer to the majority spin as "spin up" while the minority spin is "spin down." In the transition metal ferromagnets

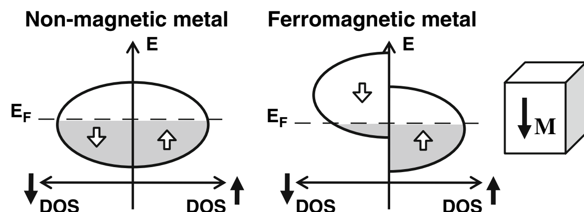


Fig. 5.1 Schematic spin-dependent density of states (DOS) for non-magnetic and ferromagnetic metals

Fe, Co, and Ni (3*d*-band ferromagnets), the orbital magnetic moment is quenched so that the magnetic moment is mostly from spin. The magnetization in Fig. 5.1 is drawn in the downward direction because the magnetic moment $\boldsymbol{m}$ is usually antiparallel to the spin ($\mathbf{m} = -g\mu_B\mathbf{S}/\hbar$, where the g-factor is usually positive and μ_B is the Bohr magneton). Because most transport properties depend on the density of states near the Fermi level, the spin asymmetry in the density of states allows ferromagnets to generate, manipulate, and detect spin.

Ferromagnetic materials also possess the property of hysteresis (Fig. 5.2), where the magnetization can have two (or more) different stable states in zero magnetic field. The bistability is due to a property called magnetic anisotropy, where the energy of a system depends on the direction of the magnetization. As shown in Fig. 5.2, there is a preferred axis ("easy axis") with stable states for magnetization direction along $\theta = 90°$ and $\theta = 270°$. When a large magnetic field ($\boldsymbol{H}$) is applied along an easy axis, the magnetization ($\boldsymbol{M}$) will align with this field in order to lower the Zeeman energy, $E_{\text{Zeeman}} = -\boldsymbol{M}\cdot\boldsymbol{H}$. When the magnetic field is turned off, the magnetization will ideally maintain all of its high-field magnetization. A magnetic field applied in the opposite direction will cause the magnetization to reverse after the field crosses a value known as the "coercivity" or "coercive field," which depends on the height of the magnetic anisotropy energy barrier (Fig. 5.2, right). This magnetic anisotropy generally depends on both the material and its shape. In terms of information storage applications, the two stable magnetic states in zero magnetic field (Fig. 5.2) correspond to the logical "0" and "1" of a data bit. The data can be written by applying a magnetic field larger than the coercivity to align the magnetization along the field. Due to the anisotropy energy barriers, this state is stable even when the magnetic field is turned off. This property makes ferromagnets natural candidates for information storage. Thus, the connection between spin and ferromagnetism establishes a natural connection between spintronics and information storage applications.

Second, spin is connected to quantum mechanics. In classical mechanics, the angular momentum (i.e., rotational motion) can be divided into two parts, an "orbital" angular momentum and a "spin" angular momentum. If one considers the motion of the earth, the elliptical motion of the earth around the sun generates orbital angular momentum, while the rotation of the earth about its

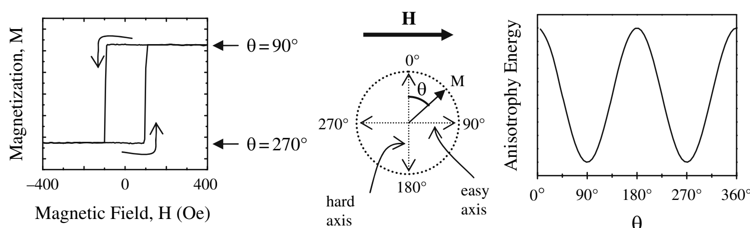


Fig. 5.2 Hysteresis loop of a ferromagnet and its origin in the magnetic anisotropy

axis generates spin angular momentum. For elementary particles such as the electron, the physics is governed by quantum mechanics. In an atom, the motion of the electron around the nucleus generates orbital angular momentum. The electron also has spin angular momentum, but it is *NOT* due to a "spinning" motion of the electron. (OK, in a dinner conversation you can describe spin as the spinning motion of an object, but technically this is wrong.) Instead, spin comes out of quantum mechanics when you try to combine it with special relativity, as Dirac did in the 1920s [26]. In solving the Dirac equation, one of the consequences is the requirement of an internal property that is now known as spin. Because of its intrinsically quantum mechanical origin, it should be of little surprise that the electron spin has very unusual properties (see Appendix). For example, its value along any particular axis (say, the z-axis) can only take on two values: $S_z = +\hbar/2$ and $-\hbar/2$. In other notations, these are called "spin up" and "spin down," $|\uparrow\rangle$ and $|\downarrow\rangle$, or $|m_s = +1/2\rangle$ and $|m_s = -1/2\rangle$. It also obeys a Heisenberg uncertainty principle where the three components of spin (S_x, S_y, S_z) cannot be measured simultaneously. Most importantly from the point of view of computing applications, the spin can be in a quantum superposition state, such as $A|\uparrow\rangle + B|\downarrow\rangle$ where the coefficients A and B are complex numbers. If you think about digital electronics as being built on bits that can have two states "0" or "1," you can think about spin as a "quantum bit" which can be in states $|\uparrow\rangle$, $|\downarrow\rangle$, or in a superposition state $A|\uparrow\rangle + B|\downarrow\rangle$, where $|A|^2$ is the probability of finding the spin in the $|\uparrow\rangle$ state and $|B|^2$ is the probability of finding the spin in the $|\downarrow\rangle$ state. The quantum bit, or "qubit," lies at the heart of a new type of proposed computer known as a quantum computer which could in principle perform some tasks such as factorizing numbers or performing searches much more efficiently than normal digital computers [27, 28, 29, 30]. There are many schemes proposed for quantum computing (with most of them being unrelated to electron spin) but there is a debate about whether a scalable quantum computer will ever be realized. Nonetheless, it is a worthy pursuit that pushes the boundaries of our knowledge and technical capabilities. Electron spin in semiconductors is a candidate for quantum information due to the long electron spin coherence times (i.e., the time over which the quantum superposition state remains well defined) observed in GaAs ($\sim$100 ns at 5 K) [31], II–VI quantum wells ($\sim$1 ns at room temperature) [32], and ZnO ($\sim$200 ps at room temperature) [33]. Furthermore, spin coherence times in materials composed of lighter elements such as silicon or carbon should be even longer due to lower spin–orbit coupling. This is several orders of magnitude larger than coherence times associated with orbital motion (i.e., phase relaxation times), so electron spin presents the opportunity to exploit quantum mechanical behavior in solids in a manner that is generally inaccessible in purely charge-based electronics. It is this special relation between spin and quantum mechanics which forms a natural connection between spintronics and advanced computing, whether the goal is a full-blown quantum computer or a more modest form of quantum information processing that has yet to be devised.

Third, the length scale of spin-dependent exchange interactions is on the order of a few atomic spacings. Because of this, the spin-dependent properties in solids are very sensitive to the atomic scale structure. With the ability to engineer interfaces at the atomic level by growth techniques such as molecular beam epitaxy (MBE) and with the ongoing improvements in the fabrication of smaller and smaller structures in the lateral dimensions, the role of spin is likely to become even more important as nanoscale science continues to advance. Understanding how spin and magnetism depend on the atomic scale interface and material structure will be an important area of investigation for the development of spintronics.

Finally, spintronics has the possibility to deliver high-speed performance and low power consumption, although one should be cautious about making such blanket statements. As electronic devices continue to shrink in size, one of the biggest problems is power dissipation and thermal management of circuits. Spin does have some potential benefits in terms of power. In terms of memory, a ferromagnetic bit can store information without any power consumption to maintain the data due to the anisotropy energy barrier (in contrast to some semiconductor memories such as SRAM or DRAM). In terms of switching, while power is required to generate electrical currents to produce magnetic fields or spin torque (Section 5.2.3), precessional dynamics can ideally proceed without dissipation (although in practice there is always at least a small amount of damping). In the case of magnetization switching by spin torque, this creates a counter-intuitive situation where the critical current required for switching a bit can be lowered by decreasing the damping parameter (e.g., through materials engineering) without decreasing the anisotropy energy barrier which stabilizes the bit [34]. More generally, novel spintronic memory or logic architectures developed in the future may be able to take advantage of precessional dynamics (of magnetization or non-equilibrium spin populations) for low-power operation. These are in principle quite advantageous, but in reality a full power analysis needs to be performed because these energy savings could be more than offset by other required operations (e.g., generating magnetic field pulses, charging up gates, etc.). Spin also has some potential benefits for speed. In charge-based electronics, the speed is set by the *RC* time constants. In utilizing spin, it may be possible to circumvent this general rule. For example, the precession of spin or magnetism (which can be at high GHz frequencies) is not governed by *RC* time constants. To sum up, there is potential for high-speed, low-power applications, but novel circuit architectures need to be developed to bring this to fruition.

5.1.2 Timeline

Before proceeding, it is instructive to display a timeline of some of the key advances to see how various lines of research have developed. Historically,

activity in the field now known as spintronics ramped up after the discovery of giant magnetoresistance (GMR) in magnetic multilayers in 1988 [23]. The research activity in these and related systems is known as “metal spintronics.” Meanwhile in the mid-1990s, the development of dilute ferromagnetic semiconductors and the discovery of long spin coherence times in semiconductors have spawned the field of semiconductor spintronics [31, 32]. More recently, there has been increased activity in lateral spin transport devices, not only in metals [35, 36] and semiconductors [37, 38] but also in newer materials such as carbon nanotubes [11, 14] and graphene [19]. In this tutorial we attempt to highlight these developments and provide an intuitive picture of the key ideas.

5.2 Metallic Magnetic Multilayers

5.2.1 Interlayer Exchange Coupling and Giant Magnetoresistance

The discoveries of interlayer exchange coupling (IEC) in Fe/Cr/Fe in 1986 [39] and giant magnetoresistance (GMR) in the same system in 1988 [23] launched the field of spintronics. For this work, Peter Grünberg and Albert Fert were awarded the 2007 Nobel Prize in Physics. While there was notable work on spin-polarized transport prior to these discoveries [35, 40, 41], GMR and IEC stimulated intense research activity not only because of the interesting physics

but also because the effects were sizeable at room temperature. Ultimately, GMR led to revolutionary advances in magnetic hard drives within 10 years of its initial discovery [42].

5.2.1.1 Interlayer Exchange Coupling

The most familiar property of ferromagnets is the magnetic force. We know that bar magnets can pick up certain metals. We also know that if you have two bar magnets, the forces can be attractive or repulsive. A "north" pole of one magnet will repel the "north" pole of another magnet, but it will be attracted to its "south" pole. These magnetic forces are due to the magnetic fields generated by the ferromagnets and these effects are known as "magnetic dipolar coupling."

In metallic multilayers consisting of alternating ferromagnetic (FM) and non-ferromagnetic (NM) layers, there is a magnetic dipolar coupling between pairs of FM layers. However, when the distance between neighboring FM layers gets small enough (i.e., less than the electron mean free path), a new type of coupling emerges that results from quantum mechanical exchange interactions. This was discovered in 1986 [39] and is known as interlayer exchange coupling (IEC).

The basic picture of IEC dates back to theoretical work in the 1950s by Ruderman and Kittel [43], Kasuya [44], and Yosida [45] (RKKY). Given two magnetic moments, $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$, inside a non-magnetic metal, they considered what happens when a conduction electron sequentially scatters off of $\boldsymbol{m}_1$ and then propagates and scatters off of $\boldsymbol{m}_2$. If the scattering depends on the spin of the electron, then it was found that the energy (E) of the system depends on the orientations of the magnetic moments as $E = -J_{\mathrm{RKKY}}\, \boldsymbol{m}_1 \cdot \boldsymbol{m}_2$. This magnetic coupling implies that the moments want to be parallel for $J_{\mathrm{RKKY}} > 0$ ("ferromagnetic coupling") and antiparallel for $J_{\mathrm{RKKY}} < 0$ ("antiferromagnetic coupling"). Interestingly, the functional form of the magnetic coupling coefficient is $J_{\mathrm{RKKY}} \sim (1/r^3)\cos(2k_{\mathrm{F}}r)$, where k_{F} is the Fermi wavevector ($k_{\mathrm{F}} = 2\pi/\lambda_{\mathrm{F}}$) of the conduction electrons and r is the distance between the two magnetic moments. This stunning theoretical result implies that the preferred magnetic orientation alternates between parallel and antiparallel as a function of their separation due to the wave nature of the electron. One way to think about this result is in terms of screening and Friedel oscillations. Due to the spin-dependent electron scattering by the magnetic moment, a cloud of spin polarization will attempt to screen the magnetic moment. Specifically, screening of the magnetic moment by an electron produces spin-polarized standing waves with the wavelength of the electron. Because there are many different electron wavelengths available, the longer wavelength oscillations from lower energy electrons get screened out by shorter wavelength oscillations from higher energy electrons. This trend continues until one reaches the Fermi energy, which is the highest energy and shortest wavelength of electrons in the metal. Because there are no electrons with shorter wavelength available to screen out the oscillations of the Fermi

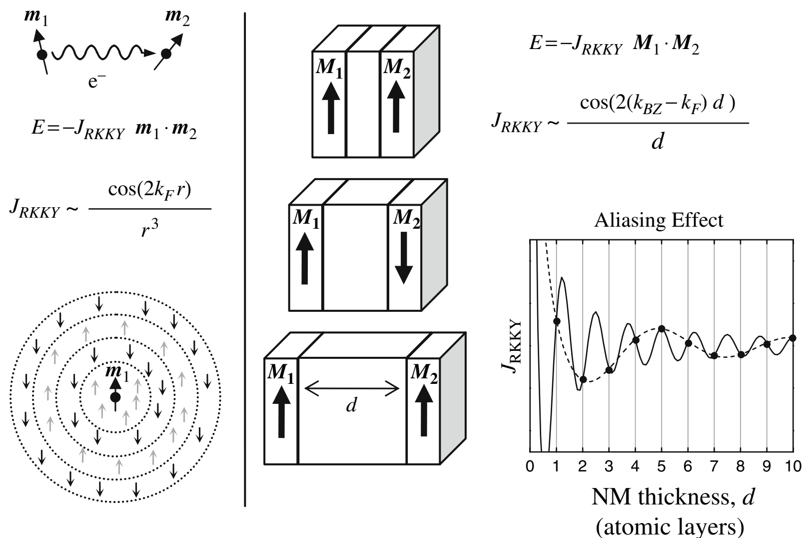


Fig. 5.3 RKKY coupling for atomic moments and multilayers

level electrons, the screening cloud will oscillate according to the Fermi wavelength λ_F as shown in Fig. 5.3 (lower left). If one next considers a second magnetic moment interacting with this electron cloud, it is clear that the coupling will oscillate with separation.

Although the RKKY theory had been useful for understanding interactions in dilute magnetic alloys such as spin glasses, the oscillatory nature of the RKKY interaction was not clearly observed until the advent of the magnetic multilayer systems in the late 1980s. Due to the very short Fermi wavelengths (approximately few angstroms), thickness control approaching atomic layer precision was needed. This was accomplished by using either sputter deposition or molecular beam epitaxy (MBE).

The IEC with antiferromagnetic coupling ($J < 0$) was first observed in 1988 in the Fe/Cr/Fe system [39]. Subsequently, the oscillatory nature of the IEC was established through systematic investigations on FM/NM/FM systems with many different NM spacer materials and thicknesses [46]. As a function of NM thickness, the magnetization of the FM layers was observed to oscillate between parallel and antiparallel alignments. In this study the IEC was observed across many different NM metal spacers, and the coupling strength oscillated as a function of spacer thickness with a period corresponding to the Fermi wavelength of the material, as predicted by the RKKY theory. In order to apply the original RKKY theory (which couples two isolated magnetic moments) to the case of a FM/NM/FM trilayer, one must sum the interaction over all pairs of magnetic moments in neighboring FM/NM interfaces, yielding a result of $J_{\mathrm{RKKY}} \sim (1/d)\cos(2k_{\mathrm{F}}d)$, where d is the NM layer thickness (solid curve in

Fig. 5.3) [47, 48]. The experimental data, however, follows the dashed curve given by $J_{RKKY} \sim (1/d)\cos(2(k_{BZ}–k_F)d)$, where $k_{BZ} = \pi/a$ is the wavevector of the Brillouin zone edge and a is the lattice constant (typically ~1–2 Å). This is due to the fact that the NM film thickness is not truly a continuous quantity because it is composed of discrete entities, namely atoms, which have a spacing given by the lattice constant a. The film thickness is often expressed in units of atomic layers (AL) and it is typical to represent this as a continuous quantity, for example 4.0, 4.5, or 5.0 AL. A fractional thickness of "0.5 AL" represents a half-filled atomic layer, so that a film with 4 AL thickness over 50% of its area and 5 AL thickness over the remaining areas is designated a "4.5 AL film." In this case, the coupling is an average of $J_{RKKY}(d = 4\ \text{AL})$ and $J_{RKKY}(d = 5\ \text{AL})$, which approximately lies on the dashed line in Fig. 5.3 given by $J_{RKKY} \sim (1/d)\cos(2(k_{BZ}–k_F)d)$. Note that although the two expressions for J_{RKKY} have noticeably different periods of oscillation, they exhibit the same values at integer multiples of atomic spacing. The key difference in the two expressions is that in deriving $J_{RKKY} \sim (1/d)\cos(2(k_{BZ}–k_F)d)$ the discrete atomic spacing, the short Fermi wavelengths, and the aliasing of the two quantities have been correctly taken into account, while in $J_{RKKY} \sim (1/d)\cos(2k_Fd)$ no such considerations were included.

More detailed studies of the oscillatory IEC were performed using wedged NM films, which were created by translating the sample behind a shutter during deposition to yield a continuous variation in film thickness across the sample (Fig. 5.4). When used in conjunction with local magnetization probes such as spin-polarized scanning electron microscopy with polarization analysis (SEMPA) or magneto-optic Kerr effect (MOKE), the systematic dependence of IEC on NM thickness was obtained [49, 50]. The main results of such studies were the systematic analysis of IEC and the identification of so-called "short period oscillations" resulting from the non-spherical nature of the Fermi surface.

While the RKKY model was highly successful in describing the dependence of IEC on the NM spacer thickness, it did not explain observations of IEC oscillations as a function of FM thickness [51]. Coupling models based on quantum well states incorporate electron wave propagation in both the NM and FM layers to account for these data, thus forming a more complete picture

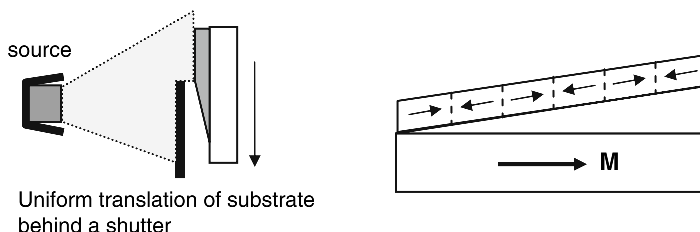


Fig. 5.4 Deposition of wedged film and systematic studies of IEC using wedges

of IEC than the RKKY model [52, 53]. Photoemission experiments have confirmed the role of quantum wells in the IEC [54, 55].

5.2.1.2 Giant Magnetoresistance

In 1988, the phenomenon of giant magnetoresistance (GMR) was discovered in Fe/Cr superlattices where the thickness of the Cr layers is such that the IEC is antiferromagnetic (Fig. 5.5) [23]. The antiparallel magnetization alignment was changed to a parallel magnetization alignment by the application of an external magnetic field (with sufficiently strong fields, the magnetizations align because the lowering of Zeeman energy overcomes the increase in IEC energy). When the in-plane resistance was measured, a change in resistance of ~50% was observed as a function of magnetic field. Given that magnetoresistance in bulk materials is typically less than a few percent, this effect was named "giant" magnetoresistance. GMR was immediately recognized for potential applications in magnetic field sensing and information storage, which are discussed in Section 5.2.4.

While sophisticated theories of GMR have been developed [56, 57, 58], the behavior can be understood qualitatively through a simple resistor model. We first point out that GMR is not due to the interaction between the conduction electrons and the magnetic field. Rather, it is due to the interaction between the conduction electrons and the FM layers via spin-dependent scattering. As such, the only role of the magnetic field is to change the relative magnetization alignment of neighboring FM layers between parallel and antiparallel.

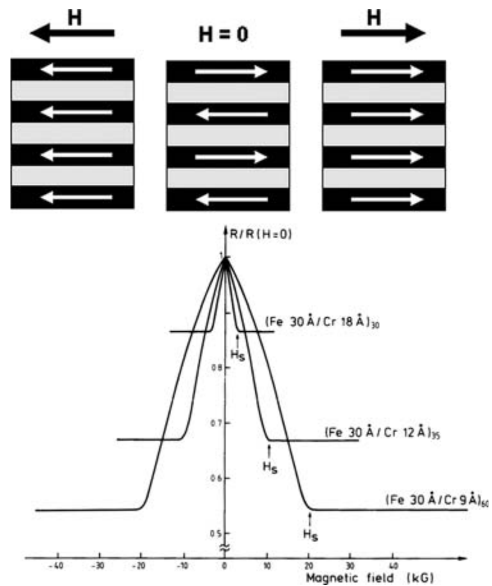


Fig. 5.5 Giant magnetoresistance in Fe/Cr superlattices. Reprinted with permission from Ref. [23]. Copyright 1988 by the American Physical Society

To model the GMR, we compare the resistance for the cases of parallel and antiparallel magnetization alignments, and for simplicity we consider just a FM/NM/FM trilayer. The current is assumed to be carried by two independent channels: a spin-up channel and a spin-down channel. We also assume that there is a spin-dependent scattering in the FM layer; for concreteness we suppose that the scattering is much stronger when the electron spin is parallel to the magnetization and much weaker when antiparallel. Finally, we assume that the layers are thinner than the mean free path of the electrons. This is important to ensure that a significant portion of the electrons will traverse both FM layers despite the fact that the average current flow is in the plane of the layers.

For the case of antiparallel magnetization alignment shown in Fig. 5.6, we first consider the spin-up channel and keep in mind that the scattering is stronger when the spin is parallel to magnetization. As shown, the spin-up electron will exhibit strong scattering in the left FM and weak scattering in the right FM (and for simplicity we ignore scattering in the NM layer). This scattering is the source of resistance and the contribution from the left FM is a large resistance, R_{large}, and the contribution from the right FM is a small resistance, R_{small}. Because electrons sample both FM layers (assuming the thicknesses are much less than the mean free path), the total resistance for the spin-up channel is obtained by adding these contributions in series: $R_{\mathrm{AP}}^{\uparrow} = R_{\mathrm{large}} + R_{\mathrm{small}}$ Similarly, for the spin-down channel the left FM contributes a small resistance, R_{small}, and the right FM contributes a large resistance, R_{large}, yielding a total resistance for the spin-down channel of $R_{\mathrm{AP}}^{\downarrow} = R_{\mathrm{small}} + R_{\mathrm{large}}$. The total resistance of the antiparallel configuration is

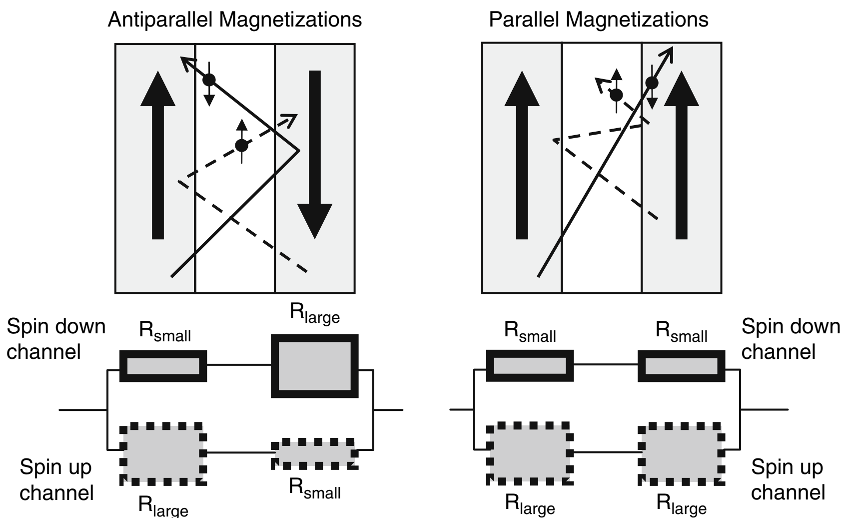


Fig. 5.6 Two channel resistance models for giant magnetoresistance

obtained by adding the resistances of the two conduction channels, which is done by adding the resistances in parallel: $R_{AP} = (R_{small} + R_{large})/2$.

For the case of parallel magnetization alignment shown in Fig. 5.6, the spin-up electron exhibits strong scattering in both the left and right FM, yielding a resistance of $R_P^{\uparrow} = R_{large} + R_{large}$. In this case, the two FM layers have "double teamed" to strongly scatter the spin-up electron. On the other hand, the spin-down electron experiences weak scattering in both FM layers, yielding a resistance of $R_P^{\downarrow} = R_{small} + R_{small}$. The total resistance of the parallel configuration is then obtained by adding the two resistances in parallel to yield $R_P = 2R_{large}R_{small}/(R_{large} + R_{small})$.

By considering an extreme limit where $R_{small} \rightarrow 0$, from the previous equations one immediately finds $R_P \rightarrow 0$, while R_{AP} remains large at $R_{large}/2$. In this case, the magnetoresistance (MR) defined as $\Delta R/R_P$ goes to infinity. Conceptually, the resistance of the parallel configuration (R_P) goes to zero for the following reason. Because the resistance of the two spin channels ($R_P^{\downarrow}$ and $R_P^{\uparrow}$) add in parallel, as the resistance of the spin-down channel ($R_P^{\downarrow} = 2R_{small}$) goes to zero due to the reduced scattering, it dominates the overall resistance to make $R_P \rightarrow 0$ (which, in turn, generates the high MR)—this behavior is often called the "short circuit effect."

If one calculates the MR without taking R_{small} to be zero, it is readily found that $\mathrm{MR} = \Delta R/R_P = (R_{AP}–R_P)/R_P = (R_{large}–R_{small})^2/4R_{large}R_{small}$. We note that without asymmetric spin scattering (i.e., if $R_{large} = R_{small}$) the MR would be zero as expected. We also note that the expression for MR is always positive, meaning that $R_{AP} > R_P$. It is possible, however, to have inverted MR if two different FM materials are chosen with opposite asymmetry of the spin-dependent scattering [59, 60].

In the above discussion, the key property for GMR is the spin-dependent asymmetry in the electron scattering. Further studies identified the critical factor as spin-dependent *interfacial* scattering, as opposed to the bulk scattering [61]. This was determined by "dusting" the interfaces in a controlled manner. For example, a few atomic layers of Co were inserted into NiFe/Cu/NiFe trilayers at different positions in the NiFe layers. If bulk scattering were most important, the MR value should not depend on the Co position within the NiFe layer. However, strong variations in MR were observed as a function of Co position, thus confirming the importance of interfacial scattering on GMR. From an engineering point of view, this is advantageous because it becomes possible to independently tune the coercivity and MR value since the former relies primarily on the bulk while the latter relies primarily on the interface.

5.2.1.3 Spin Valves

The term "spin valve" was coined to describe a FM/NM/FM trilayer device operating on the GMR principle. In subsequent years, it has become a more general term to describe a spin transport device with two FM electrodes that exhibits different resistances for parallel and antiparallel magnetization

alignments. To date, the spin valve effect has been observed in metallic multilayers, magnetic tunnel junctions, carbon nanotubes, ultrathin graphite, inorganic semiconductors, and organic semiconductors. While the detailed mechanism for the resistance change is not the same in all cases, they all originate from the spin of the electron.

To heuristically describe the spin valve effect, it is often useful to employ an analogy with optics. It is well known that light can be linearly polarized using a polarizer. By sending light through two polarizers in series, the final intensity depends on the relative alignment of the two polarizers. As shown in Fig. 5.7, if the two polarizers are parallel, then the light output is maximized. If the two polarizers are perpendicular, then the light output goes to zero. The reason is that the first polarizer will allow only the vertical polarization to pass, while the second will allow only the horizontal polarization to pass. The two polarizers work together to block all the light. The spin valve is similar, with spin polarization playing the role of light polarization and ferromagnets playing the role of the optical polarizers.

In the case of GMR, the general idea of the optical analogy still holds, but the details do not agree. For example, the current does not flow from one FM layer to the other, but instead flows parallel to the films. In addition, no net spin polarization is accumulated between the two FM layers, in contrast to the optical experiment which has linearly polarized light in between the two polarizers. The optical analogy holds more strongly for vertical transport, where the current flows perpendicular to the layers. Such a geometry is called "current-perpendicular-to-the-plane" and has been realized experimentally and studied theoretically [62, 63]. Vertical transport is also employed in magnetic tunnel junctions, which are discussed next.

5.2.2 *Magnetic Tunnel Junctions*

A device very similar to the optical analogy of a spin valve in Fig. 5.7 is the magnetic tunnel junction (MTJ), which consists of two ferromagnetic layers (F_1, F_2) separated by an insulating tunnel barrier. Due to the spin-polarized density of states, there is a high conductance when the magnetizations are parallel and a low conductance when the magnetizations are antiparallel. Figure 5.8 illustrates the effect, where for simplicity we assume 100% spin polarization at the Fermi level (later, we calculate Julliere's formula which does not assume this). In the parallel configuration, the spin-up electrons from F_1 tunnel into the spin-up states in F_2, so there is a high conductance. In

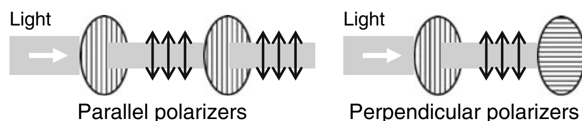


Fig. 5.7 Optical analogy of a spin valve

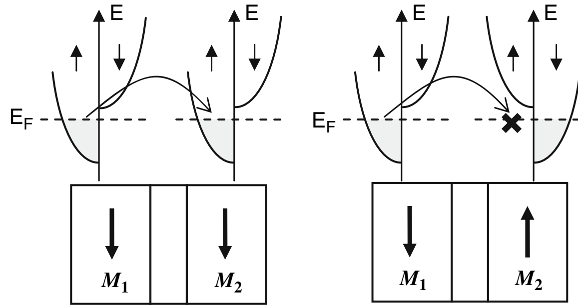


Fig. 5.8 Schematic picture of tunneling magnetoresistance

the antiparallel configuration, the spin-up electrons from F_1 cannot tunnel efficiently because there are no spin-up states in F_2, so there is a low conductance. The difference in resistance between the two configurations is known as tunneling magnetoresistance (TMR).

TMR was first observed at low temperatures in Fe/oxidized Ge/Co tunnel junctions by Julliere in 1974 [41]. Subsequently, very little work was performed on the subject until the mid-1990s when Moodera [24] and Miyazaki [64] independently observed room temperature TMR in MTJ consisting of an Al_2O_3 tunnel barrier. Optimization of the MTJ increased the values of TMR from initial values of ~10% up to values of ~70% by the late 1990s.

A schematic hysteresis loop of TMR for a Co/Al_2O_3/CoFe tunnel junction is shown in Fig. 5.9. Unlike the case of GMR, there is little IEC to help achieve an antiparallel alignment. Typically, two different FM materials with different coercivities are used to achieve the antiparallel alignment. Beginning at negative field, both magnetizations are along the negative direction. As the field is swept up (solid lines) the Co magnetization switches first to achieve an antiparallel alignment. At this point the resistance increases by the spin-dependent

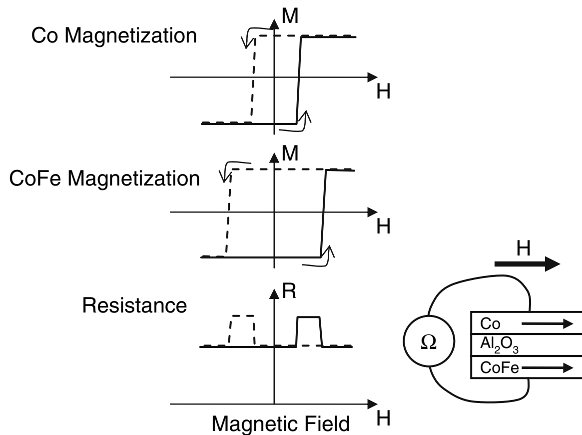


Fig. 5.9 Schematic hysteresis loop of a magnetic tunnel junction

tunneling effect discussed in Fig. 5.8. At a higher field, the CoFe magnetization switches to achieve a parallel alignment again, and the resistance switches back down to a lower value. A down-sweep of the field (dashed line) will have similar features except that the switching occurs at the negative fields.

5.2.2.1 Julliere Model for TMR

A simple expression for TMR was derived by Julliere [41] (Fig. 5.10). We perform this derivation here. The two ferromagnetic layers with magnetizations M_1 and M_2 can be made of different materials. The assumptions in the calculation are that there is no spin-dependent scattering and the tunneling rate is proportional to the product of the initial and final density of states at the Fermi level (valid for low bias voltages). We define D_i and d_i as the Fermi level density of states of the majority and minority electrons, respectively, of the ith FM layer ($i = 1, 2$). The Fermi level spin polarization of the ith ferromagnet is then defined as $P_i = \frac{D_i - d_i}{D_i + d_i}$. Taking $\Delta_i = D_i + d_i$ as the total density of states, one can write

$$D_i = \frac{\Delta_i}{2}(1 + P_i) \text{ and } d_i = \frac{\Delta_i}{2}(1 - P_i) \tag{5.1}$$

Now we calculate the TMR given by

$$\mathrm{TMR} \equiv \frac{\Delta R}{R_\mathrm{P}} = \frac{R_\mathrm{AP} - R_\mathrm{P}}{R_\mathrm{P}} = \frac{R_\mathrm{AP}}{R_\mathrm{P}} - 1 = \frac{G_\mathrm{P}}{G_\mathrm{AP}} - 1 \tag{5.2}$$

where R_AP (G_AP) is the resistance (conductance) of the antiparallel magnetization configuration and R_P (G_P) is the resistance (conductance) of the parallel magnetization configuration.

For the parallel configuration, the total conductance is the sum of the conductances of the two spin channels: $G_\mathrm{P} = G_\mathrm{P}^\uparrow + G_\mathrm{P}^\downarrow$. The conductance of each channel is proportional to the tunneling rate which is assumed to be proportional to the product of the initial and final density of states:

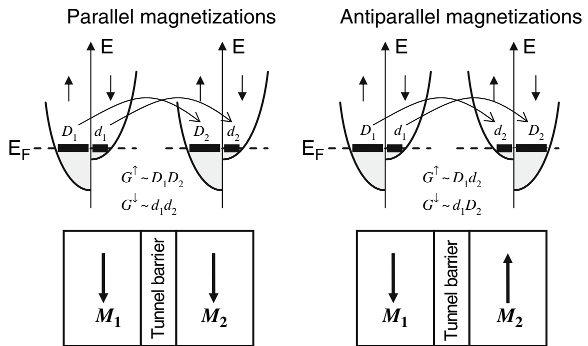


Fig. 5.10 Julliere model for tunneling magnetoresistance

$G_P^\uparrow = A\ D_1 D_2$ and $G_P^\downarrow = A\ d_1 d_2$, where A is the common proportionality constant. For the antiparallel configuration, we similarly obtain $G_{AP}^\uparrow = A\ D_1 d_2$ and $G_{AP}^\downarrow = A\ d_1 D_2$. Putting this together, we get

$$\mathrm{TMR} = \frac{G_P}{G_{AP}} - 1 = \frac{G_P^\uparrow + G_P^\downarrow}{G_{AP}^\uparrow + G_{AP}^\downarrow} - 1 = \frac{D_1 D_2 + d_1 d_2}{D_1 d_2 + d_1 D_2} - 1 \tag{5.3}$$

Using Equation 5.1 to substitute for D_1, D_2, d_1, and d_2 and performing some algebra, one obtains the well-known expression for TMR:

$$\mathrm{TMR} = \frac{2P_1 P_2}{1 - P_1 P_2} \tag{5.4}$$

The Julliere model serves as a starting point for analyzing MTJs, but there are some important limitations of this model. First, the TMR depends only on properties of the ferromagnetic layers and incorporates none of the properties of the tunnel barrier (e.g., barrier height, thickness). Second, there is no distinction between the different types of electrons within the ferromagnet, such as *s*-band electrons and *d*-band electrons, which can have different polarizations and different effective masses.

5.2.2.2 MgO-Based Magnetic Tunnel Junctions

A fantastic failure of the Julliere model is for the case of MgO-based MTJs, which were developed beginning in the early 2000s [65, 66, 67, 68] with most of the early work on Fe/MgO/Fe(001). The observed TMR values were found to far exceed those predicted by the Julliere model. Currently, the record for room temperature TMR is 410% in MBE-grown Fe/Co/MgO/Co/Fe MTJs [69]. The origin of enhanced TMR is a novel spin-filtering effect based on wavefunction symmetry [70, 71, 72], which applies to Fe/MgO(001) and bcc Co/MgO(001) junctions.

The main difference between the Al_2O_3 and the MgO tunnel barrier is that the former is polycrystalline while the latter is single crystalline or highly textured along (001). This single crystal nature of MgO, combined with some special properties of Fe (001), leads to an enhancement of spin polarization of the tunneling electrons. The limitation of Julliere's model is that the tunneling rates of all electrons are assumed to be equal, but this is not true for MgO. In the bandgap of single crystal MgO, tunneling is governed by transport through evanescent states which are characterized by the symmetry of their Bloch states (i.e., electron wavefunction). If the atomic orbitals making up the Bloch states are spherically symmetric (i.e., originating from *s*-orbitals), then the tunneling rates are significantly higher than for other states. These special states with high tunneling rates are said to have "Δ_1 symmetry." For Fe along the (001) direction, the majority spin possesses states with Δ_1 symmetry at the Fermi level, as well as states with other symmetries. However, the minority spin does *not* have

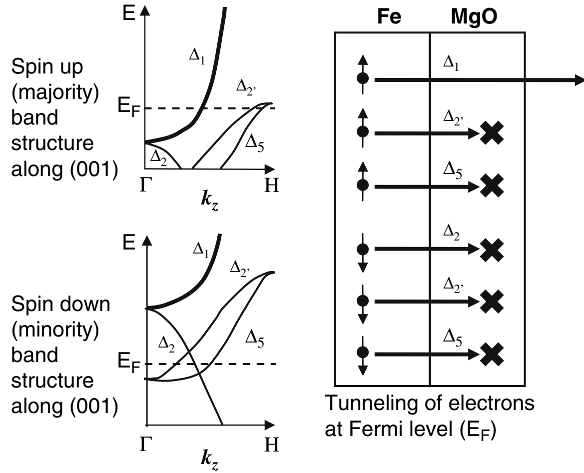


Fig. 5.11 Spin filtering of Δ_1 states by MgO for enhanced TMR

states of Δ_1 symmetry at the Fermi level. This is a very fortunate situation that is summed up in Fig. 5.11. We take the initial spin polarization in the FM to be given by $P = (N_\uparrow - N_\downarrow)/(N_\uparrow + N_\downarrow)$, where $N_\uparrow$ and $N_\downarrow$ are the total number of spin-up and spin-down electrons at the Fermi level, respectively. As they tunnel across the MgO barrier, only electrons from the Δ_1 bands of Fe can couple to the Δ_1 bands of the MgO for a high tunneling rate. All other electrons have a low tunneling rate, as indicated by the $\times$ in Fig. 5.11. Thus, a larger fraction of the spin-up electrons get across the barrier, and the spin polarization of the tunneling electrons is significantly larger than the spin polarization of the FM material itself. This enhancement of spin polarization is the origin of the high TMR values that have been observed.

This Δ_1 spin filtering also applies to the case of bcc Co ferromagnets because the band structure along (001) is similar to Fe (bcc) in terms of the Δ_1 bands. Because Co is usually hexagonal, the bcc phase is stabilized by growth on either Fe or MgO. The highest TMR of 410% is observed in Fe/Co/MgO/Co/Fe MTJs [69], while the highest value for Fe/MgO/Fe is 180% [68]. The difference may be due to differences in the ideal band structures of the junctions or due to the fact that Co is more resistant to oxidation than Fe. Further studies are needed to determine the cause of higher TMR in Co. Finally, we note that theory predicts TMR values in thousands of percent [71, 73], so further significant improvements may be possible.

5.2.3 *Spin Torque*

Both the GMR and TMR effects discussed above have a similar character, namely that the electron transport properties are strongly affected by the

magnetic configuration of a device. An interesting question to ask is whether an inverse process can occur: can the magnetic configuration be affected by electron transport? In 1996, Berger [74] and Slonczewski [75] studied this question theoretically and predicted the presence of a "spin torque" resulting from spin-dependent reflection/transmission at FM/NM interfaces and angular momentum conservation.

The spin torque was later observed experimentally in two different types of studies. In one experiment, a GMR multilayer was contacted by a point contact and peaks in dI/dV as a function of applied magnetic field were attributed to the excitation of spin waves by spin torque [76]. In another experiment, a GMR trilayer was contacted by a metal through a nanopore (Fig. 5.12) [25]. In measuring the current–voltage characteristic of the device, sharp changes in resistance corresponding to magnetization switching were observed. We follow the sign convention of Ref. [25], where a positive current corresponds to electron flow from the thinner FM layer (free layer) to the thicker FM layer (fixed reference layer). If the two FM layers are composed of the same material, the thicker layer resists switching because it has a larger magnetic anisotropy energy barrier (which scales with volume). As shown in Fig. 5.12, when a large enough positive current is applied, the resistance jumps into a high state. When a large enough negative current is applied, the resistance jumps into a low state. From our knowledge of GMR, we know that the high resistance corresponds to the antiparallel magnetization alignment whereas the low resistance corresponds to the parallel alignment. Thus, a positive current generates antiparallel alignment while a negative current generates parallel alignment. The key feature of both of these studies is the high current density generated by the nanocontacts.

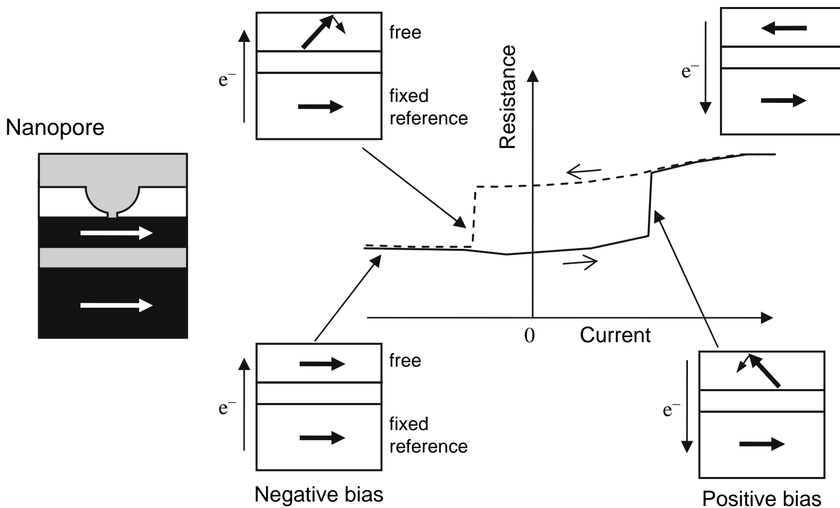


Fig. 5.12 Magnetization reversal by spin torque

To understand the origin of spin torque and the role of precessional dynamics, we first develop a simplified picture of spin torque ignoring the magnetization dynamics given by the Landau–Lifshitz–Gilbert (LLG) equation. Then we incorporate the LLG equation to understand the role of precessional dynamics.

5.2.3.1 Origin of Spin Torque

The primary source of spin torque is a spin-filtering effect whose origin is rooted in quantum mechanics [74, 75]. While other mechanisms contributing to spin torque have been identified [77], we will not discuss them here.

Figure 5.13 shows the device geometry for our calculation. $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are the net spin angular momenta of the localized magnetic moments in the two ferromagnetic layers F_1 (fixed reference layer) and F_2 (free layer). Here, we discuss the magnetization in terms of its angular momentum $\boldsymbol{S}$ (which is typically antiparallel to the magnetization) because the mechanism for spin torque is based on angular momentum conservation. $\boldsymbol{S}_1$ is fixed along the $+z'$ axis of the lab frame ($x'y'z'$), while $\boldsymbol{S}_2$ initially lies in the x'–z' plane with an angle θ away from the z' axis. A second set of coordinates (xyz) is defined such that the z-axis is parallel to $\boldsymbol{S}_2$. The vectors $\hat{i}, \hat{j}, \hat{k}$ are the unit vectors for the xyz frame, while $\hat{i}', \hat{j}', \hat{k}'$ are the unit vectors for the $x'y'z'$ frame. The spin of the free

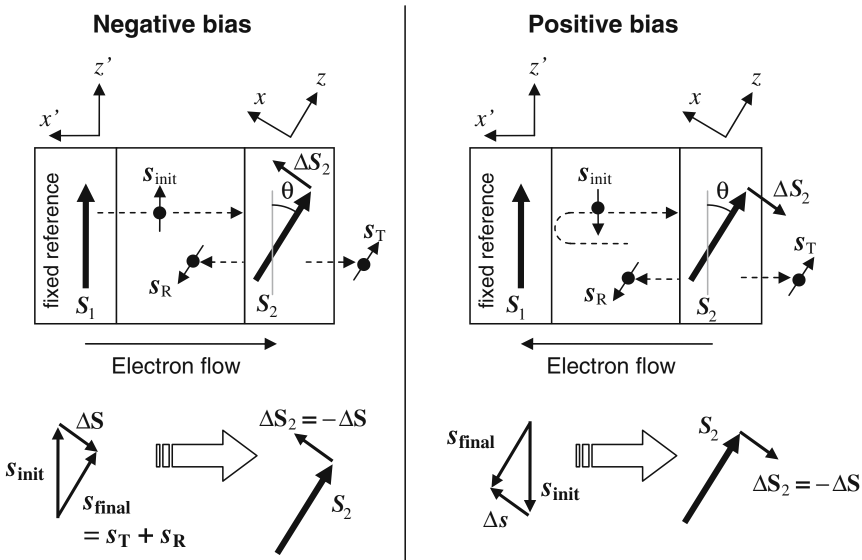


Fig. 5.13 Derivation of spin torque based on spin-dependent reflection/transmission and conservation of angular momentum. Negative bias favors parallel alignment, while positive bias favors antiparallel alignment

electrons is given by $\boldsymbol{s}$ (lowercase) and must be treated quantum mechanically (the spin operator $\tilde{\boldsymbol{s}}$ is written with a tilde and the expectation values are written without a tilde). For simplicity, we assume that the electron transmission and reflection coefficients exhibit complete spin asymmetry: electrons with spin parallel to $\boldsymbol{S}_i$ (i = 1, 2) are completely transmitted, while electrons with spin antiparallel to $\boldsymbol{S}_i$ are completely reflected. In real systems where complete spin asymmetry is not achieved, the final result is reduced by a spin-transfer efficiency parameter ($\eta \leq 1$), which includes the effects of incident spin polarization and scattering efficiency.

We first consider the case where electrons flow from the reference layer (F_1) to the free layer (F_2) (i.e., negative bias). The spin of the electrons transmitted from the reference layer is aligned with the $+z'$ direction: $\chi_{\text{init}} = |\uparrow\rangle_{z'}$, where the subscript z' indicates the quantization axis. When the electrons encounter the free layer, the electron transmission and reflection must be calculated using the $\boldsymbol{S}_2$ direction (i.e., z-axis) as the quantization axis: $\chi_{\text{init}} = |\uparrow\rangle_{z'} = \cos(\theta/2)|\uparrow\rangle_z + \sin(\theta/2)|\downarrow\rangle_z$ (see Equations 5.26 and 5.27 with $\phi = 0°$). For perfect spin asymmetry of the transmission and reflection coefficients, the first term is fully transmitted and the second term is fully reflected: $\chi_{\text{T}} = \cos(\theta/2)|\uparrow\rangle_z$ and $\chi_{\text{R}} = \sin(\theta/2)|\downarrow\rangle_z$. The expectation values for the spin operator $\tilde{\boldsymbol{s}} = \tilde{s}_x\hat{i} + \tilde{s}_y\hat{j} + \tilde{s}_z\hat{k} = \tilde{s}'_x\hat{i}' + \tilde{s}'_y\hat{j}' + \tilde{s}'_z\hat{k}'$ for these states are given by

$$\mathbf{s}_{\text{init}} = {}_{z'}\langle\uparrow|\tilde{\boldsymbol{s}}|\uparrow\rangle_{z'} = \frac{\hbar}{2}\hat{k}' = \frac{\hbar}{2}(\cos\theta)\hat{k} + \frac{\hbar}{2}(\sin\theta)\hat{i} \tag{5.5}$$

$$\boldsymbol{s_T} = {}_z\langle\uparrow|\cos(\theta/2)\ \tilde{\boldsymbol{s}}\ \cos(\theta/2)|\uparrow\rangle_z = \frac{\hbar}{2}(\cos^2(\theta/2))\hat{k} \tag{5.6}$$

$$\boldsymbol{s_R} = {}_z\langle\downarrow|\sin(\theta/2)\ \tilde{\boldsymbol{s}}\ \sin(\theta/2)|\downarrow\rangle_z = -\frac{\hbar}{2}(\sin^2(\theta/2))\hat{k} \tag{5.7}$$

The final spin is $\boldsymbol{s}_{\text{final}} = \boldsymbol{s}_{\text{T}} + \boldsymbol{s}_{\text{R}} = \frac{\hbar}{2}(\cos^2(\theta/2) - \sin^2(\theta/2))\hat{k} = \frac{\hbar}{2}(\cos\theta)\hat{k}$, and the change in spin is thus $\Delta\mathbf{s} = \boldsymbol{s}_{\text{final}} - \boldsymbol{s}_{\text{init}} = -\frac{\hbar}{2}(\sin\theta)\hat{i}$. We note that this difference is due to the presence of cross terms, $\langle\chi_{\text{R}}|\tilde{s}|\chi_{\text{T}}\rangle$ and $\langle\chi_{\text{T}}|\tilde{s}|\chi_{\text{R}}\rangle$, in the initial state expectation value (if one calculates in the xyz frame). Because the total angular momentum ($\boldsymbol{s} + \boldsymbol{S}_2$) must be conserved during this process, $\Delta\boldsymbol{s} + \Delta\boldsymbol{S}_2 = 0$, or equivalently $\Delta\boldsymbol{S}_2 = -\Delta\mathbf{s} = +\frac{\hbar}{2}(\sin\theta)\hat{i}$. This is the angular momentum change of F_2 per incident electron, assuming perfect spin-transfer efficiency ($\eta = 1$). The analysis is summarized by the vector diagrams in the lower part of Fig. 5.13. A first conclusion is that under negative bias, the spin torque on $\boldsymbol{S}_2$ will cause the free layer magnetization to align *parallel* with the fixed reference layer magnetization. A second conclusion is that the spin torque is maximized for $\theta = 90°$. When there is a current of electrons, the spin torque on F_2 is $N_2 = \frac{dS_2}{dt} =$ (# electrons per unit time)($\Delta\boldsymbol{S}_2$ per electron) $= \left(\frac{|I|}{e}\right)\frac{\hbar}{2}\eta(\sin\theta)\hat{i}$ where I is the

current, e is the magnitude of electron charge, and η is the spin-transfer efficiency (which depends on the incident spin polarization, scattering efficiency, etc.) If one defines $\hat{s}_1$ and $\hat{s}_2$ as unit vectors along $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$, then the spin torque on F_2 can be written as $\boldsymbol{N}_2 = -\left(\frac{|I|}{e}\right)\frac{\hbar}{2}\eta\hat{s}_2 \times (\hat{s}_2 \times \hat{s}_1)$.

We next consider the case where electrons flow from the free layer to the reference layer (i.e., positive bias). The analysis is similar to the previous case with one important difference. With electrons flowing toward the fixed reference layer, the spin of reflected electrons will point along the $-z$ direction (under our assumption of perfect spin asymmetry). We take this as the initial spin: $\chi_{\text{init}} = |\downarrow\rangle_{z'}$. These reflected electrons will subsequently interact with the free layer. The analysis follows as before, but the opposite sign of $\boldsymbol{s}_{\text{init}}$ leads to an opposite sign for the spin torque. The situation is summarized by the vector diagrams in the lower part of Fig. 5.13. Thus, under positive bias, the spin torque on $\boldsymbol{S}_2$ will cause the free layer magnetization to align *antiparallel* with the fixed reference layer magnetization. In this case, the spin torque on F_2 is $\boldsymbol{N}_2 = \left(\frac{|I|}{e}\right)\frac{\hbar}{2}\eta\hat{s}_2 \times (\hat{s}_2 \times \hat{s}_1)$. Combining the expressions for positive and negative bias, the spin torque is $\boldsymbol{N}_2 = \frac{d\boldsymbol{S}_2}{dt} = \left(\frac{I}{e}\right)\frac{\hbar}{2}\eta\hat{s}_2 \times (\hat{s}_2 \times \hat{s}_1)$, where the sign of I indicates the bias.

The key result of the spin torque analysis is simply that *positive bias favors antiparallel alignment and negative bias favors parallel alignment*. In the next section, we will consider the role of magnetization dynamics, where it is customary to discuss the ferromagnetic layers in terms of their magnetizations $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ instead of $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$: $\boldsymbol{M}_1 = -g\mu_B\boldsymbol{S}_1/\hbar V_1$, $\boldsymbol{M}_2 = -g\mu_B\boldsymbol{S}_2/\hbar V_2$, where V_1 and V_2 are the volumes of the layers. (More properly, we should also include the orbital magnetic moment as well.) Although it is usually cumbersome to keep track of the negative sign between $\boldsymbol{M}$ and $\boldsymbol{S}$, the key result for spin torque is the same for the $\boldsymbol{M}$s: positive bias favors antiparallel magnetization alignment and negative bias favors parallel magnetization alignment of $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ (assuming the same sign for the g-factors).

5.2.3.2 Excitation of Precessional Dynamics

As mentioned earlier, it was found that the spin torque could excite the precessional magnetization dynamics [76, 78]. With precession frequencies in the microwave regime, this potentially enables new types of microwave sources and detectors. To understand how this occurs, we first need to understand standard magnetization dynamics, which is governed by the Landau–Lifshitz–Gilbert (LLG) equation. In general, the dynamics of the magnetization $\boldsymbol{M}$ is described by the LLG equation as

$$\frac{d\boldsymbol{M}}{dt} = -\gamma\boldsymbol{M} \times \boldsymbol{H}_{\text{tot}} + \frac{\alpha}{M_S}\boldsymbol{M} \times \left(\frac{d\boldsymbol{M}}{dt}\right) \tag{5.8}$$

where γ is the gyromagnetic ratio, α is the Gilbert damping parameter, M_S is the magnitude of $\boldsymbol{M}$, and $\boldsymbol{H}_{tot}$ is the total magnetic field which includes the external field and internal fields (these include the demagnetization and anisotropy fields). For magnetic films, the demagnetization field is perpendicular to the film and has a value of $-4\pi M_\perp$, but for simplicity we shall ignore the internal fields and take $\boldsymbol{H}_{tot} = \boldsymbol{H}$. An equivalent second form of the LLG equation is

$$\frac{\mathrm{d}\boldsymbol{M}}{\mathrm{d}t} = -\gamma'\boldsymbol{M} \times \boldsymbol{H}_{tot} - \frac{\gamma'\alpha}{M_S}\boldsymbol{M} \times (\boldsymbol{M} \times \boldsymbol{H}_{tot}) \tag{5.9}$$

where $\gamma' = \gamma/(1+\alpha^2)$. Typically $\alpha << 1$, so that $\gamma' \approx \gamma$. The first term describes the precessional motion, while the second term describes the damping.

To gain intuition about the LLG equation, one should consider a situation in which a large magnetic field $\boldsymbol{H}$ is turned on instantaneously in a direction perpendicular to $\boldsymbol{M}$, as shown in Fig. 5.14. In order to lower the Zeeman energy ($E_{Zeeman} = -\boldsymbol{M}\cdot\boldsymbol{H}$), the magnetization will want to align with $\boldsymbol{H}$. However, $\boldsymbol{M}$ will not follow a direct path. Instead, it follows a spiraling path that eventually ends up aligning with $\boldsymbol{H}$ (Fig. 5.14). This motion can be understood by looking at the vectorial directions of the two terms on the right-hand side of the LLG equation. The precession term $-\gamma'\boldsymbol{M} \times \boldsymbol{H}$ is always perpendicular to $\boldsymbol{M}$ and $\boldsymbol{H}$ so the resulting dynamics is a circular precession. The damping term $-\frac{\gamma'\alpha}{M_S}\boldsymbol{M} \times (\boldsymbol{M} \times \boldsymbol{H})$ generates a short vector that points toward $\boldsymbol{H}$, which is responsible for the spiraling behavior and the ultimate alignment of $\boldsymbol{M}$ with $\boldsymbol{H}$.

We now discuss the impact of spin torque on the LLG equation. The interplay of spin torque and precessional dynamics can be investigated by adding the spin torque term to the LLG equation for the free layer magnetization $\boldsymbol{M}_2$:

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{M}_2}{\mathrm{d}t} =& -\gamma'\boldsymbol{M}_2 \times \boldsymbol{H}_{tot} - \frac{\gamma'\alpha}{|M_2|}\boldsymbol{M}_2 \times (\boldsymbol{M}_2 \times \boldsymbol{H}_{tot}) \\ &+ \frac{g_2\mu_B J\eta}{2ed_2}\left(\frac{g_1}{|g_1|}\right)\hat{m}_2 \times (\hat{m}_2 \times \hat{m}_1)\end{aligned} \tag{5.10}$$

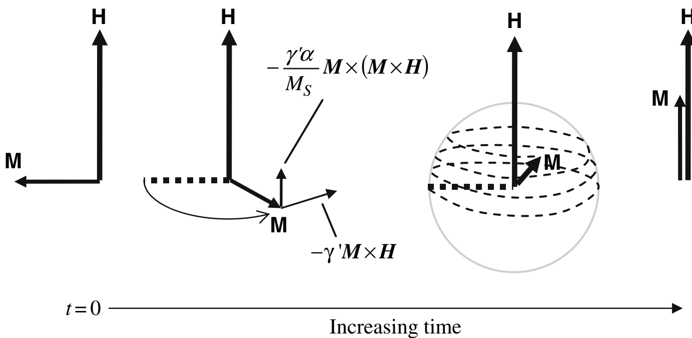


Fig. 5.14 Landau-Lifshitz-Gilbert magnetization dynamics

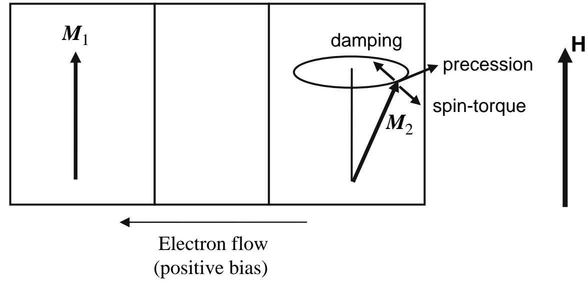


Fig. 5.15 Steady-state magnetization precession is achieved when the spin torque cancels the damping

where g_1 and g_2 are the g-factors of the two FM layers, d_2 is the thickness of F_2, J is the current density, and $\hat{m}_1 = \frac{\boldsymbol{M}_1}{|\boldsymbol{M}_1|}$ and $\hat{m}_2 = \frac{\boldsymbol{M}_2}{|\boldsymbol{M}_2|}$.

We consider a situation in which $\boldsymbol{M}_1$ (fixed layer) and $\boldsymbol{H}$ lie along the z-axis. In this case, the last two terms—the damping and spin torque terms—have the same functional form. Thus, if the current density J is positive (i.e., positive bias), then the spin torque can counteract the effects of damping and pull $\boldsymbol{M}_2$ away from the z-axis. Figure 5.15 shows the situation where the damping and spin torque cancel to yield a steady-state precession of the magnetization. This accounts for the steady precession in an applied magnetic field. When one includes the internal fields (anisotropy and demagnetization fields) the motion is more complex, but the general idea still holds.

In zero magnetic field, the spin torque just generates a magnetization switching. However, the magnetization does not follow a direct path, as suggested in Fig. 5.12. Due to the presence of internal anisotropy fields and demagnetization fields ($-4\pi M_{\perp}$ for thin films), even in zero external field the magnetization switching follows a spiral path instead of a direct path [33].

Further studies on spin torque have addressed the roles of magnetic domain structure and time-resolved dynamics [79, 80]. Phase-locked coupled oscillations driven by spin torque have also been observed [81, 82]. Spin torque is now a rather widespread phenomenon observed in a variety of contexts. In addition to metallic magnetic multilayers, spin torque has been observed in MTJs [83, 84], FM nanowires [85, 86, 87, 88, 89, 90], and lateral spin valves with non-local spin injection [91].

5.2.4 Applications

Magnetic multilayer devices have become very important for information storage technologies. In order to understand the use of magnetic multilayers and MTJs for storage applications, it is important to understand the phenomenon of exchange bias, discovered in 1956 [92, 93]. We consider the magnetic properties of a ferromagnet/antiferromagnet bilayer, shown in Fig. 5.16. An antiferromagnet is a material with magnetic ordering, but with a net magnetization of zero. The antiferromagnet shown in Fig. 5.16 consists of magnetic

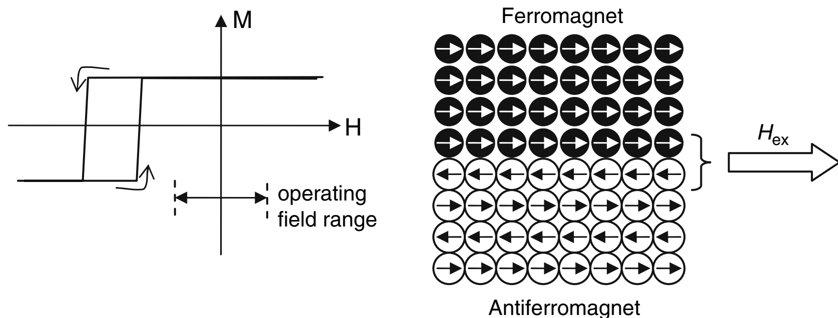


Fig. 5.16 The exchange bias effect in ferromagnet/antiferromagnet bilayers with exchange field on the ferromagnetic layer, H_{ex}, originating from the interface

moments which alternate direction with each atomic plane. Because the magnetization is zero, the antiferromagnet is relatively insensitive to an applied magnetic field. Technically speaking, there is some response to magnetic fields since the moments are able to tilt laterally, but we will ignore such effects for this discussion. The topmost atomic layer of the antiferromagnet provides an exchange coupling which biases the magnetic switching properties of the adjacent FM layer. Usually the interfacial energy prefers that the magnetic moments at the FM/AF interface are antiparallel, so the topmost AF layer can be modeled as providing an internal "exchange field" H_{ex} on the FM layer pointing to the right. The presence of this additional field causes the magnetic hysteresis loop of the FM layer to shift to the left, as shown in Fig. 5.16, and this phenomenon is known as exchange bias. The key point is that in the operating magnetic field range of this device, the FM magnetization is always positive (to the right). Thus, the effect of the AF layer is to "pin" the FM layer to always point to the right. This could be used, for example, to pin the reference layer in the spin torque structures discussed in the previous section. We note that this discussion of exchange bias is highly simplified and more detailed discussions are available in recent review articles [94, 95].

5.2.4.1 Magnetic Hard Drives

The GMR effect has been utilized as magnetic field sensors in the read heads of magnetic hard disk drives (Fig. 5.17). Within 10 years of the initial discovery of GMR, the first GMR hard drives were developed [42]. GMR is largely responsible for the great increases in hard drive capacity from the late 1990s to the present. This has played an important role in the emergence of the Internet and digital video. The GMR technology has become the dominant technology for hard drives, and a natural evolution is to take advantage of the high TMR values in MTJs.

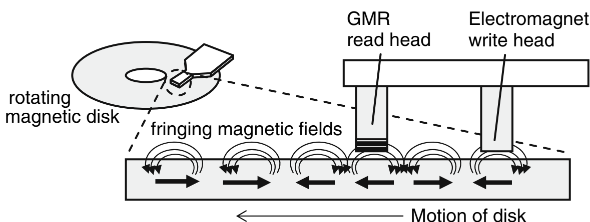


Fig. 5.17 Key elements of a magnetic hard drive

Magnetic hard drives store information on a magnetic disk, which spins at high speed. Each data bit corresponds to the magnetization of one region of the disk, and the data are written using a small electromagnet that produces a magnetic field larger than the coercivity of the magnetic material on the disk. Once written, these magnetic regions produce "fringing" magnetic fields. The reading process involves detecting the direction of these fringing magnetic fields. While many elements are needed for a complete disk drive system (e.g., control system, lubrication, etc.), our discussion completely centers on the magnetic field sensor of the read head, which has proven to be one of the most important elements for high-density storage.

For the read head, the GMR effect is used as a magnetic field sensor. The basic metallic multilayer structure is shown in Fig. 5.18. The top three layers are a FM/NM/FM trilayer which exhibits MR due to the GMR effect. An antiferromagnetic layer is adjacent to the bottom FM layer. Through the exchange bias effect the bottom FM layer is pinned to always point to the right. For a sensitive magnetic field sensor, a material having a low coercivity is chosen for the top FM layer ("free layer"). This ensures that the magnetization direction will track the applied magnetic field. A typical candidate for the free layer is permalloy ($Ni_{81}Fe_{19}$), which has a very low coercivity of about 1 Oersted (similar to the Earth's magnetic field). When the magnetic field points to the right, the top FM layer magnetization will point to the right as well, resulting in a low resistance due to the parallel magnetization alignment of the two FM layers. When the magnetic field points to the left, the top FM layer magnetization will also point to the left, resulting in an antiparallel magnetization alignment and a high resistance.

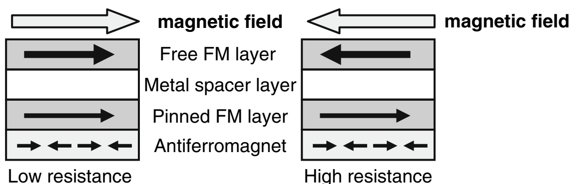


Fig. 5.18 GMR spin valve as a magnetic field sensor

5.2.4.2 Magnetic Random Access Memory (MRAM)

MTJs are being utilized for non-volatile solid-state memory known as magnetic random access memory (MRAM) [96, 97, 98, 99]. While it is still unclear whether MRAM will become a dominant technology for non-volatile memory, two key advantages are the high speed and durability. The write speed is limited by magnetization precession dynamics, and write times of a few ns have been demonstrated (compared to μs writing times for Flash). The durability comes from the fact that changing memory states does not require high voltages and does not involve the motion of atoms.

MRAM is based on an array of MTJs as shown in Fig. 5.19. Each MTJ stores one bit of data, and the data are addressed by a unique pair of electrodes (bit line, word line). Of the two FM layers in the MTJ, the bottom FM layer in Fig. 5.19 is adjacent to an antiferromagnetic layer and is pinned to always point to the right. The top FM electrode is free to switch. The magnetization of this layer is what stores the data. For concreteness, we say that the logical "0" corresponds to the free layer magnetization pointing to the right, while logical "1" corresponds to the free layer magnetization pointing to the left.

Reading a bit is performed by measuring the resistance. Due to the TMR effect, the "0" state will have a low resistance and the "1" state will have a high resistance. Writing the bit is more difficult. Commercial MRAM chips perform the writing by using localized magnetic fields, while the next generation technology is expected to use spin torque to write the data.

The first method to write an MRAM bit is to use localized magnetic field pulses generated by current pulses in a "half-select" approach. To address a particular bit, one must choose a particular horizontal wire and vertical wire as shown in Fig. 5.20. Along each of these two wires, a current pulse is applied to generate a magnetic field along the wire (via Biot–Savart law). The value of this

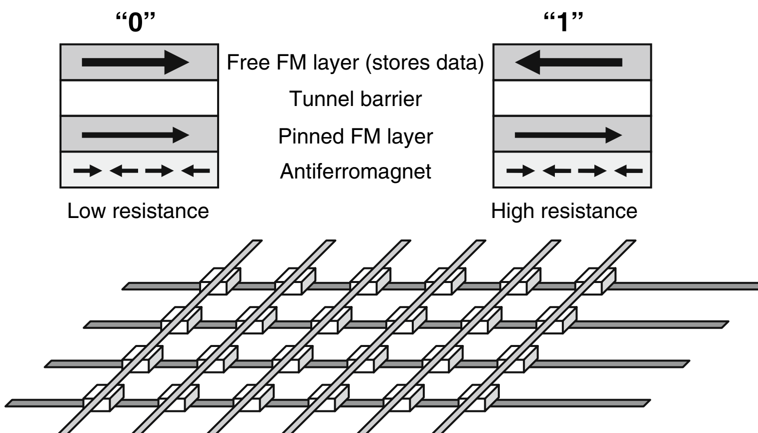


Fig. 5.19 MRAM array based on MTJ memory elements

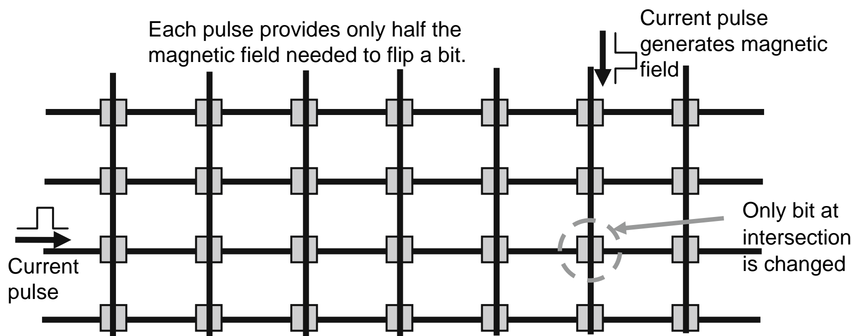


Fig. 5.20 Half-select approach to writing bits in MRAM

magnetic field is critical. A serious problem is encountered if the magnetic field from a single wire is larger than the coercivity. All the bits along the wire will switch—a horrible result. Therefore the magnetic field must be less than the coercivity of a magnetic bit, so that the bits along the wire are not switched. At the intersection of two wires the magnetic field is the sum of the fields generated by each wire. This allows the bit at the intersection to be switched without affecting other bits. Because each wire only provides half of the required field for switching, this method is called "half-select."

In this half-select approach, a "0" is written by applying current pulses to generate a magnetic field in the "right" direction, as shown in Fig. 5.19, and a "1" is written by applying current pulses in the opposite direction. While this method is straightforward, the selectivity is not good enough. Due to the difficulty in confining the localized magnetic fields, neighboring bits can be accidentally switched. Furthermore, this crosstalk problem becomes worse if the bit density is increased by bringing the MTJs closer together. A clever solution to this problem is "toggle-MRAM," which is used in commercial MRAM today [97]. The basic idea is that the free FM layer is replaced by an antiferromagnetically coupled FM/NM/FM metallic trilayer. The "0" and "1" states are shown in Fig. 5.21. High or low resistance values are obtained because the TMR effect is sensitive only to the FM layers adjacent to the tunnel barrier. In this scheme, a current pulse induces a canting of the FM layers of the AF-coupled trilayer and causes magnetizations to flip. If the initial state is "0," then

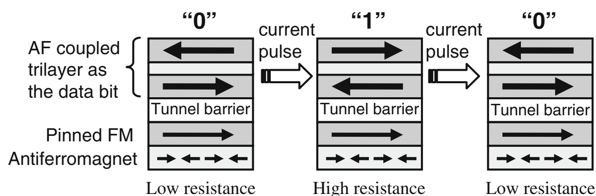


Fig. 5.21 Memory element for toggle-MRAM

a current pulse will switch it to "1." If the initial state is "1," then the current pulse will switch it to "0." This toggling action of the current pulse means that to write a bit, one must first read the bit and then send a current pulse only if necessary. While this writing scheme is more complicated, the big advantage of toggle-MRAM is that the sensitivity of the AF-coupled trilayer to stray magnetic fields drops rapidly with distance. This dramatically suppresses the crosstalk problems mentioned earlier and has made it possible to commercialize MRAM.

An alternative method for writing MRAM bits is spin torque. By applying a large enough current density, the "0" and "1" states can be written using the proper polarity of the current. Because spin torque switching has been observed in MTJs, this method is viable for MRAM [83, 84]. The main advantage is the suppression of crosstalk because the switching mechanism is confined to the bit that is being addressed. However, a number of technical challenges need to be addressed, such as reducing the critical current density for switching and developing MTJs that combine high TMR with low resistance-area product. Nonetheless, the use of spin torque provides a promising avenue to increase the MRAM bit density.

5.3 Semiconductor Spintronics

Semiconductors form the backbone of electronics and computing as a result of their highly tunable transport properties. By changing the impurity doping level, the carrier density can be tuned over several orders of magnitude, and the charge of the carriers can be either negative (electrons) or positive (holes). In addition, electrostatic gates can significantly adjust the carrier density in real time. These properties are utilized in bipolar junction transistors and field effect transistors which are used to amplify analog signals or perform digital logic operations. Semiconductor heterostructures routinely employ bandgap engineering (using the dependence of bandgap on composition or strain) to tailor the potential energy landscape experienced by the carriers. Finally, some semiconductors have excellent optoelectronic properties for generating or detecting light.

Due to such capabilities of semiconductors, there has been great interest in coupling to the spin degree of freedom. Progress has been made in a number of areas. Incorporating ferromagnetism into a semiconductor via dilute magnetic doping has yielded new, tunable magnetic behavior [100, 101, 102, 103]. Optical studies have demonstrated extremely long spin coherence times [31], opening up new possibilities for quantum information processing in the solid state. Finally, the generation of spin polarization in semiconductors from ferromagnets has been achieved using spin injection [104, 105] and spin reflection [106, 107]. These establish building blocks for lateral semiconductor spin devices, which are discussed more in Section 5.4.

5.3.1 Ferromagnetic Semiconductors

A dilute magnetic semiconductor (DMS) is a semiconductor alloy in which a small concentration of magnetic ions (typically less than 10%) is introduced into a non-magnetic semiconductor material [100]. Ferromagnetism was first demonstrated in the III–V DMS (In,Mn)As [108] and (Ga,Mn)As [109]. These systems exhibit ferromagnetism only at temperatures below 200 K. Before any technological applications can be realized, an increase in the ferromagnetic ordering temperature (T_C) is necessary. To this end, there are many efforts to explore ferromagnetism in other DMS systems, including (but not limited to) magnetically doped Ge [110, 111] and magnetically doped ZnO and GaN [112]. However, because the III–V DMSs are currently the most well-studied and characterized ferromagnetic semiconductors, we will limit our discussion to these systems.

(Ga,Mn)As is synthesized by low temperature (100–300°C) MBE with a Mn concentration typically between 0 and 10% for a homogeneous alloy [100]. At high Mn concentration and/or high growth temperature, secondary phases such as ferromagnetic MnAs or GaMn clusters begin to form. For homogeneous alloys, the magnetic ordering temperature (T_C) depends on Mn concentration and on the growth procedures. As grown, T_C is typically below 100 K but can increase to $\sim$150 K upon low-temperature annealing [113, 114].

The origin of ferromagnetism in semiconductors is a topic of active interest. Similar to the RKKY coupling discussed earlier, two Mn magnetic moments can be coupled indirectly through a free carrier—in this case a hole. The important role of holes in mediating the magnetic coupling between spatially separated Mn moments was first seen experimentally through two different observations: (1) the correlation of T_C with the hole concentration in (Ga,Mn)As [115] and (2) the photoinduced ferromagnetism in (In,Mn)As [103]. Because the spacing between Mn atoms is shorter than the Fermi wavelength (due to low hole density compared to metals), the oscillations in the RKKY coupling are not realized and the earlier Zener model [116] is sufficient. Dietl et al. applied the Zener model and calculated several properties of (Ga,Mn)As including T_C, magnetic anisotropy, and magnetic circular dichroism [117, 118]. In addition, this model predicted high T_C in magnetically doped GaN and ZnO, which stimulated a large search for high-T_C ferromagnetism in DMS. In a heuristic description of this model, the holes are coupled antiferromagnetically to the Mn. When two Mn moments interact with the same hole, the Mn moments will both prefer to align antiparallel to the hole's moment, so they will prefer to align parallel with each other. Thus, the lowest energy is achieved when all the Mn moments are parallel with each other and antiparallel with the holes, as shown in Fig. 5.22.

Further studies have identified some important factors for ferromagnetism in (Ga,Mn)As. Studies of low-temperature annealing found that T_C is increased substantially ($\sim$150 K) when the sample is annealed near the growth

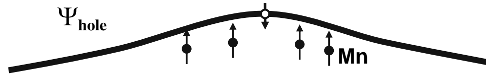


Fig. 5.22 Carrier-mediated ferromagnetism in dilute magnetic semiconductors. Interactions between localized Mn moments and extended hole wavefunctions lead to parallel alignment of Mn

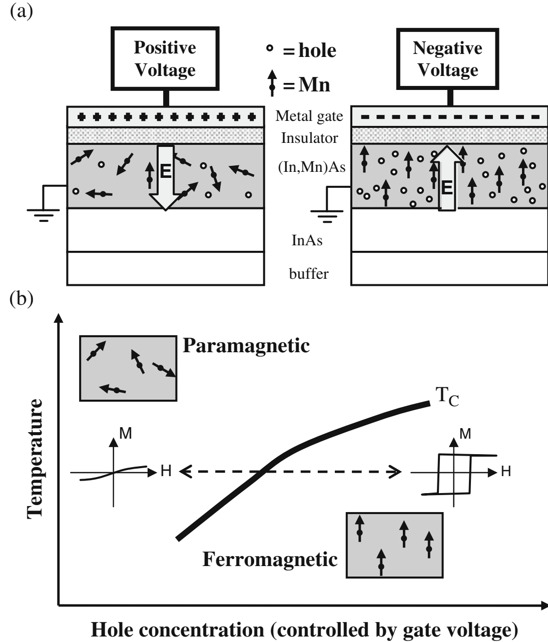
temperature (∼100–300°C) for up to several hours [113, 114]. Rutherford backscattering measurements find that this T_C enhancement occurs because Mn moves from an interstitial site to a Ga-site [119]. This implies that Mn on Ga-sites is actively participating in the ferromagnetism, while interstitial Mn does not. To understand the role of interstitial Mn, theoretical studies have been performed [120]. Optical magnetic circular dichroism studies also find a coupling between the Mn moments and the valence band edge spectra (i.e., holes) [121]. Recent optical spectroscopy measurements have identified the formation of a Mn impurity band, which may play a very important role in the formation of ferromagnetic ordering [122]. Alternative methods for introducing Mn includes delta-doping, which in some cases has led to high T_C (∼170 K) [123, 124].

While the carrier-mediated ferromagnetism is still not completely understood, some interesting devices have already been realized. Specifically, the electric field control of ferromagnetism was first demonstrated in (In,Mn)As [101]. Using a field effect transistor structure shown in Fig. 5.23a, the ferromagnetism was turned on and off reversibly by applying a gate voltage. This can be understood through Fig. 5.23b. The effect of the gate voltage is to change the hole concentration at fixed temperature (dashed arrow), which in turn increases the T_C. This allows the system to reversibly change from a disordered state (paramagnetic) to an ordered state (ferromagnetic), and vice versa, by controlling the gate voltage. This effect has been observed in (In,Mn)As (∼20 K) [101], (Ga,Mn)As (∼60 K) [125], II–VI DMS ($<$ 10 K) [102], and Mn delta-doping in GaAs (∼110 K) [126]. Tunable ferromagnetism may become useful for applications if the operating temperatures are increased.

5.3.2 *Optical Studies of Spin Coherence*

Ultrafast optical techniques for investigating electron spin dynamics in semiconductors were developed in the 1990s [127]. This effort led to the discovery of long spin coherence times in semiconductors (∼150 ns in GaAs [31]) and the ability to transport spin over macroscopic distances (∼100 μm) [128]. These techniques also enabled studies on the manipulation of spin by a variety of means including spin–orbit effects [129], *g*-factor engineering [130], optical fields [131], and ferromagnets [106, 107].

Fig. 5.23 Electric field control of carrier-mediated ferromagnetism



The primary techniques for measuring spin dynamics in semiconductors are time-resolved Faraday rotation (TRFR) and time-resolved Kerr rotation (TRKR), where the former is a transmission measurement while the latter is a reflection measurement. In the Faraday (Kerr) effect, a linearly polarized optical beam is transmitted through (reflected from) a spin population, causing the polarization axis to rotate by an angle proportional to the spin-polarization component along the beam path.

The TRFR measurement of spin dynamics in a direct gap semiconductor such as GaAs relies on short (~150 fs) pulses generated from a Ti:sapphire laser. The pulses are generated at a high repetition rate (76 MHz) but for the moment let us consider just a single pulse. The beamsplitter (BS) in Fig. 5.24b splits this pulse into two separate pulses. One pulse acts as a "pump" while the other acts as a "probe." The TRFR measurement sequence is shown in Fig. 5.24a. The pump pulse is circularly polarized and arrives at the GaAs sample first. The wavelength of this pulse is tuned to the bandgap of GaAs and the absorption of circularly polarized light leads to the generation of spin-polarized electrons with spin oriented perpendicular to the sample [132]. After a time delay Δt, a linearly polarized probe pulse arrives at the GaAs sample to measure the spin polarization via Faraday rotation (i.e., rotation of polarization axis). The rotation angle (θ_F) is proportional to the component of spin polarization along the beam path (S_x). The spin dynamics are obtained by performing this measurement for different values of Δt (Δt is usually stepped through a range of values).

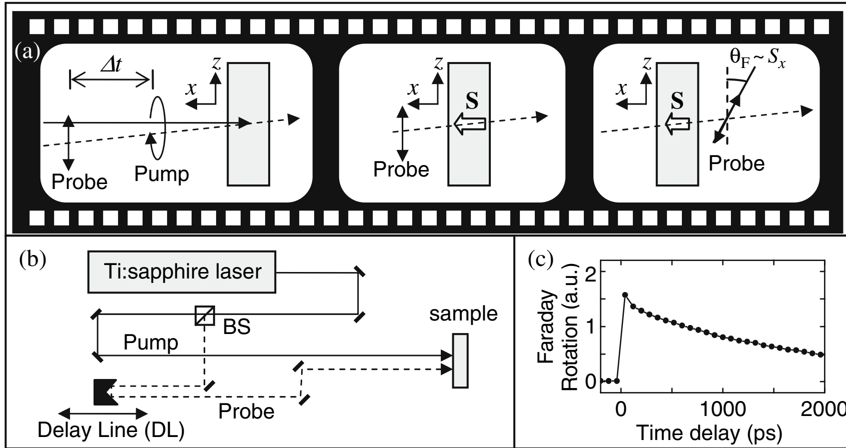


Fig. 5.24 (**a**) A filmstrip of time-resolved Faraday rotation: a circularly polarized pump pulse generates spin in the semiconductor, and a time-delayed, linearly polarized probe pulse measures the spin after a time Δt. (**b**) Adjustment of the time delay (Δt) by a mechanical delay line. (**c**) Decay of spin polarization in zero magnetic field as a function of Δt

An interesting part of this measurement is that the time delay Δt is controlled by a mechanical delay line (DL). Because the speed of light is $c = 3.0 \times 10^8$ m/s, moving the mirrors in the DL by a distance of 1 mm changes Δt by $2 \times 1\,\text{mm}/c = 6.7\,\text{ps}$. Thus, by rather modest position control of the DL, very good temporal resolution can be achieved (ultimately limited by the duration of the laser pulse). To measure the dynamics of the spin excited by the pump beam, the DL is stepped through a range of positions corresponding to a range of Δt values and the Faraday rotation is collected at each position. Figure 5.24c shows data of Faraday rotation vs. Δt, for spin excitation and decay in zero magnetic field.

In practice, the measurement is usually not based on a single pulse, but a train of pulses repeated at 13 ns interval (76 MHz). The measured signal is therefore the result of 76 million experiments per second, leading to a high signal-to-noise ratio. If the spin lifetime is shorter than 13 ns, then the spin excitations are independent and the measured signal represents the average dynamics from a single excitation. On the other hand, if the spin lifetime exceeds 13 ns, subsequent pulses interfere and the more complex behavior of resonant spin amplification is observed [31].

To investigate the dephasing of the photoexcited spin population, the same measurement is performed with a magnetic field applied along the z-axis in Fig. 5.24a. As we will show, the spins will precess about the magnetic field (similar to the magnetization precession discussed in Section 5.2.3). The magnetic field defines the quantization axis for the spin. Quantum mechanically, the

initial spin along the x-axis is equal to a superposition of the spin-up and spin-down states along the z-axis: $|\uparrow\rangle_x = \frac{|\uparrow\rangle_z + |\downarrow\rangle_z}{\sqrt{2}}$ (see Equation 5.28). In the presence of a magnetic field, the energy levels of the spin states will split according to the Zeeman effect: $\Delta E = E_\uparrow - E_\downarrow = g\mu_B H$, where g is the g-factor, μ_B is the Bohr magneton, and H is the magnetic field along the z-axis. From the Schrödinger equation, the time evolution of a quantum eigenstate is known to be $\phi(t) = \exp(-iE_\phi t/\hbar)\phi(0)$. Taking the initial state of the spins as $\chi(0) = |\uparrow\rangle_x = \frac{|\uparrow\rangle_z + |\downarrow\rangle_z}{\sqrt{2}}$, the Schrödinger time evolution yields

$$
\begin{aligned}
\chi(t) &= \frac{\exp(-iE_\uparrow t/\hbar)|\uparrow\rangle_z + \exp(-iE_\downarrow t/\hbar)|\downarrow\rangle_z}{\sqrt{2}} \\
&= \exp(-iE_\uparrow t/\hbar)\frac{|\uparrow\rangle_z + \exp(i\Delta E t/\hbar)|\downarrow\rangle_z}{\sqrt{2}} \\
&= \exp(-iE_\uparrow t/\hbar)\frac{|\uparrow\rangle_z + \exp(ig\mu_B H t/\hbar)|\downarrow\rangle_z}{\sqrt{2}}
\end{aligned}
\tag{5.11}
$$

Due to the energy splitting of the states, a relative phase accumulates between the two states leading to a quantum beating of the states. The physical interpretation of this beating is obtained by calculating the expectation value of the spin operator (see Equation 5.27 with $\theta = 90^\circ, \phi = g\mu_B H t/\hbar$):

$$S_x(t) = \langle \tilde{S}_x \rangle = \left\langle \chi(t) \left| \frac{\tilde{S}_+ + \tilde{S}_-}{2} \right| \chi(t) \right\rangle = \frac{\hbar}{2}\cos\left(\frac{g\mu_B H t}{\hbar}\right) \tag{5.12}$$

$$S_y(t) = \langle \tilde{S}_y \rangle = \left\langle \chi(t) \left| \frac{\tilde{S}_+ - \tilde{S}_-}{2i} \right| \chi(t) \right\rangle = \frac{\hbar}{2}\sin\left(\frac{g\mu_B H t}{\hbar}\right) \tag{5.13}$$

$$S_z(t) = \langle \tilde{S}_z \rangle = \langle \chi(t)|\tilde{S}_z|\chi(t)\rangle = 0 \tag{5.14}$$

The expectation value of spin is a vector which precesses about the z-axis with frequency $\omega_L = g\mu_B H/\hbar$.

Experimental measurement of the spin dynamics in a transverse magnetic field indeed shows oscillations in S_x (Fig. 5.25) [31]. In addition, the amplitude exhibits an exponential decay which is associated with dephasing and decoherence of the spin population. The data are fit by the curve:

$$S_x(t) = S_0 \exp\left(-\frac{t}{T_2^*}\right)\cos\left(\frac{g\mu_B H t}{\hbar}\right) \tag{5.15}$$

where T_2^* is the transverse spin lifetime. This provides a lower bound on the spin coherence time, which represents the loss of fidelity of the quantum spin state.

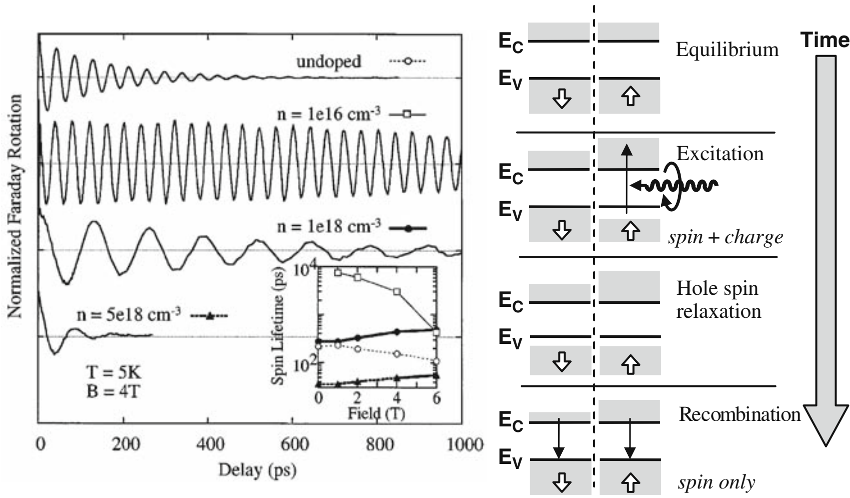


Fig. 5.25 *(Left)* Ultrafast optical measurement of electron spin precession in GaAs for different levels of n-type doping concentrations. Reprinted with permission from Ref. [31]. Copyright 1998 by the American Physical Society. (*Right*) A schematic time sequence to achieve the pure spin excitation needed for long spin lifetimes

Experiments performed on bulk n-type GaAs wafers produced surprising results. The transverse spin lifetime exhibited strong dependence on bulk doping density, with an optimal doping in the 10^{16} cm^{-3} range. Near this excitation density, the transverse spin lifetime was found to be as high as $\sim$100 ns at 5 K. The doping dependence of spin lifetime is understood as follows. For undoped samples, the spin lifetime is limited by the exciton recombination time because when all electrons finish recombining with holes, there are no carriers left. Without carriers, there can be no spin polarization. For doped systems, electrons are present in equilibrium. Therefore, spin excitations can persist even after electron–hole recombination has been completed. For low doping ($\sim 10^{16}$ cm^{-3}), this leads to a very long spin lifetime. As shown schematically in Fig. 5.25, circularly polarized light generates a population of spin-polarized excitons. The spin polarization of photoexcited carriers is actually 50% due to optical selection rules [132], but drawn as 100% for simplicity. This is followed by rapid depolarization of the hole spins. The subsequent electron–hole recombination leaves a spin polarization in the conduction band while the valence band is absent of holes. Now there is a spin excitation without charge excitation, so the spin lifetime is limited only by spin-dependent interactions. At higher doping, the spin lifetime is reduced due to spin–orbit coupling, which is discussed below. This type of long spin lifetime has been observed at room temperature in II–VI quantum wells [32], GaN [133], and ZnO [33].

5.3.2.1 Role of Spin–Orbit Coupling

The role of spin–orbit coupling is extremely important for spins in semiconductors, providing both desirable and undesirable properties. For one, the spin lifetime in bulk semiconductors is limited by the spin–orbit coupling, so lower spin–orbit coupling is desired. On the other hand, spin–orbit coupling is needed to optically generate and detect spins. Furthermore, spin–orbit coupling could be used to generate or manipulate spins. We will discuss the physical idea of spin–orbit coupling and some of its effects on spin.

Spin–orbit coupling is a relativistic effect that arises when you consider the same situation from two different reference frames. In the lab frame (Fig. 5.26, left), consider an electron flying past a positively charged nucleus at rest. The nucleus creates an electric field but no magnetic field (hence no coupling to spin) because it is not moving. On the other hand, in the electron's frame (Fig. 5.26, right) the nucleus is moving. This motion of positive charge generates a current which produces a magnetic field by the usual Biot–Savart law. This field, which we denote $\boldsymbol{H}_{\mathrm{SO}}$, interacts with the spin's magnetic moment ($\boldsymbol{m}$) via a standard Zeeman energy term ($E_{\mathrm{Zeeman}} = -\boldsymbol{m} \cdot \boldsymbol{H}_{\mathrm{SO}}$). Because of its relativistic origin, the internal magnetic field $\boldsymbol{H}_{\mathrm{SO}}$ increases with the velocity of the electron.

More generally, when an electric field $\boldsymbol{E}$ is present in the lab frame, an internal magnetic field $\boldsymbol{H}_{\mathrm{SO}}$ is generated in the electron's reference frame. The form of this field is known as the Rashba spin–orbit coupling and is described by $\boldsymbol{H}_{\mathrm{SO}} \sim \boldsymbol{E} \times \boldsymbol{k}$, where $\boldsymbol{k}$ is the wavevector (i.e., momentum) of the electron [134]. In a zinc-blende solid, which does not possess inversion symmetry, there is another type of internal spin–orbit field known as the Dresselhaus spin–orbit field. Based on a symmetry analysis, it is found that the form of this internal field is $\boldsymbol{H}_{\mathrm{SO}} \sim k_x\left(k_y^2 - k_z^2\right)\hat{i} + k_y\left(k_z^2 - k_x^2\right)\hat{j} + k_z\left(k_x^2 - k_y^2\right)\hat{k}$ [132, 135]. A key feature apparent in both types of spin–orbit coupling is that the effective field $\boldsymbol{H}_{\mathrm{SO}}$ depends on the momentum of the electron. Given a population of spin-polarized electrons, there is a distribution of momenta and therefore a distribution of $\boldsymbol{H}_{\mathrm{SO}}$. Thus each electron will experience precession along a different internal field axis, causing the net spin polarization of the population to decay. This dephasing mechanism is known as Elliot–Yafet when the spin–orbit coupling originates from impurities and Dyakonov–Perel when the spin–orbit coupling is generated intrinsically from the ideal band structure of the material [132].

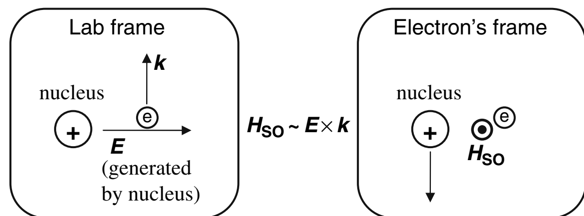


Fig. 5.26 Spin–orbit coupling originating from a change of reference frames

In the experiment shown in Fig. 5.25, spin lifetime decreases with increased doping (above 10^{16} cm^{-3}) due to the spin–orbit coupling. At higher doping levels, more of the conduction band is filled and the wavevectors of the electrons get bigger. This leads to stronger spin–orbit effects, and consequently an increased dephasing rate and shorter spin lifetimes.

Although the dephasing introduced by spin–orbit coupling is not desired, an attractive aspect of spin–orbit coupling is that it could be used to manipulate spins. If a population of spins is moving with some average drift velocity, there is a non-zero average wavevector $\langle \boldsymbol{k} \rangle$. Considering the Rashba coupling, this generates a non-zero average internal field $\langle \boldsymbol{H}_{\mathrm{SO}} \rangle \sim \boldsymbol{E} \times \langle k \rangle$ when an electric field is present. The Zeeman splitting and spin precession resulting from this internal field is known as the "Rashba effect" [134]. In 1990, Datta and Das proposed a spin transistor based on the Rashba effect [136]. The Datta–Das spin transistor consists of ferromagnetic source and drain electrodes which inject and detect spin in a two-dimensional electron gas (2 DEG) channel (Fig. 5.27). This idea is similar to the spin valves discussed earlier. The main difference, however, is the presence of an electrostatic gate which can generate an electric field perpendicular to the 2 DEG ($\boldsymbol{E}$ along the z-axis). As the electrons flow from the source to the drain ($\langle k \rangle$ along the x-axis), the Rashba effect generates an average internal field $\langle \boldsymbol{H}_{\mathrm{SO}} \rangle$oriented along the y-axis. Because the spin orientation is perpendicular to $\langle \boldsymbol{H}_{\mathrm{SO}} \rangle$, it will experience a precession about the y-axis, with a frequency that depends on the electric field controlled by the gate voltage as $\omega_{\mathrm{L}} = g\mu_{\mathrm{B}}|\langle \boldsymbol{H}_{\mathrm{SO}} \rangle|/\hbar \sim g\mu_{\mathrm{B}}|\langle \boldsymbol{k} \rangle||\boldsymbol{E}|/\hbar$. When the spins reach the drain electrode, the relative orientation of the spin with the drain magnetization will determine the source–drain current, with maximum current for parallel and minimum current for antiparallel alignment. Because the final orientation of the spin depends on the precession frequency, which in turn depends on the gate voltage, the dependence of source–drain current on the gate voltage should look something like the curve shown in Fig. 5.27. Interesting aspects of this device are that small changes in gate voltage could lead to sharp changes in source–drain current, and a negative differential transconductance can be achieved.

The proposal of the Datta–Das spin transistor provided motivation to develop semiconductor spintronic devices and was far ahead of its time. At present, this type of spin transistor has yet to be realized, and it has taken many years to demonstrate the basic building blocks required for this device. The process of spin injection into a semiconductor was not convincingly

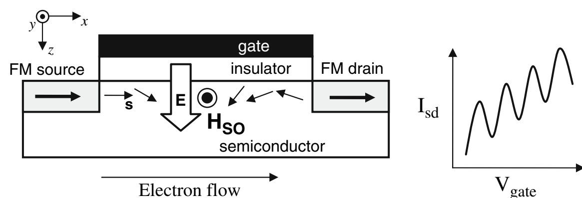


Fig. 5.27 *(Left)* Schematic drawing of the Datta-Das spin transistor. (*Right*) The oscillatory behavior of source-drain current with gate voltage due to spin precession

demonstrated until 1999 [104, 105] and is discussed further in the next section. The Rashba effect was observed through optical experiments in 2004 [129] and is described next. Finally, all-electrical injection and detection of spin in semiconductors was achieved only recently [37, 38]. With these various ingredients coming together, there is a chance to develop novel lateral spin transport devices, which will be the topic of Section 5.4.

Returning to the Rashba effect, an ultrafast optical measurement provided the first direct demonstration of this effect [129]. Instead of using ferromagnets to inject and detect the spin, this experiment used optical pulses to inject spin (circularly polarized pump) and to detect spin (Faraday rotation of linearly polarized probe). In this experiment, a lateral bias was applied to a GaAs film, while the pump and probe spots were located at different points of the film, as shown in Fig. 5.28. Instead of utilizing an electric field normal to the plane of the film, a strain gradient was applied. This was achieved by two methods: (1) using the natural bend of a free-standing GaAs membrane and (2) employing strained InGaAs films on GaAs substrates. The resulting strain effectively changes the bandgap, so a strain gradient along the z-direction produces a potential energy gradient (∇V) for the conduction electrons along the z-direction. This potential energy gradient (∇V) plays the role of the electric field and produces an internal magnetic field given by $\langle \boldsymbol{H}_{\mathrm{SO}} \rangle \sim \nabla V \times \langle \boldsymbol{k} \rangle$. We note that the standard relations among the electric field ($\boldsymbol{E}$), electrical potential (Φ), and potential energy (V) for electrons are: $\boldsymbol{E} = -\nabla\Phi$, $V = (-e)\Phi$, $\boldsymbol{E} = (1/e)\nabla V$. With electron flow, $\langle \boldsymbol{k} \rangle$, along the x-axis, the internal field is along the y-axis (Fig. 5.25). A circularly polarized pump pulse generates spin polarization along the z-axis. As the spins are dragged laterally along the x-axis, they precess about the internal field $\langle \boldsymbol{H}_{\mathrm{SO}} \rangle$, as shown in Fig. 5.28. A linearly polarized probe pulse then detects the z-component of spin polarization. By scanning the probe position to measure the spin polarization as a function of position and time, the spin precession due to the Rashba effect was clearly observed. We note that a similar spin-orbit-induced spin precession was achieved by applying a uniaxial stress to GaAs [137].

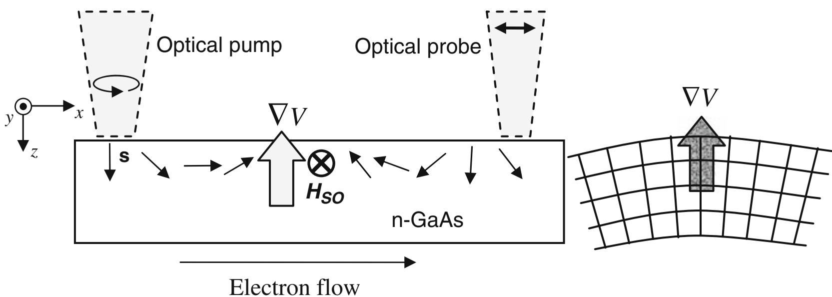


Fig. 5.28 Optical measurement of the Rashba effect in n-GaAs produced by strain gradients

5.3.3 Ferromagnet|Semiconductor Structures: Spin Injection and Accumulation

5.3.3.1 Spin Injection: The Spin-LED

The basic process of injecting spin-polarized electrons from a ferromagnet into a non-magnetic semiconductor proved to be a major challenge. While current-perpendicular-to-the-plane (CPP) GMR studies clearly demonstrated spin injection from a ferromagnet into a non-magnetic metal in all-metal structures [62], a clear demonstration of spin injection into semiconductors proved to be much more difficult. In hindsight, the main difficulty was the conductivity mismatch between the metallic ferromagnet and the semiconductor, which was not well appreciated until after spin injection was achieved in all-semiconductor structures.

Spin-dependent light-emitting diode (spin-LED) experiments provided the first definitive demonstration of electrical spin injection into semiconductors. Instead of using a ferromagnetic metal as the spin injector, these studies utilized either a ferromagnetic semiconductor (GaMnAs) or a paramagnetic semiconductor (BeMnZnSe) as the spin injector [104, 105]. An integrated p–i–n LED structure provided a means for detecting the spin polarization optically. Describing the (Ga,Mn)As experiment, a voltage bias is applied to inject spin-polarized holes from the (Ga,Mn)As (p-type) into the GaAs (intrinsic). These holes then recombine in an InGaAs quantum well with unpolarized electrons from an n-type GaAs injector. As shown in Fig. 5.29, the helicity of the emitted light depends on the spin polarization of the injected spin-polarized holes. By measuring the circular polarization of the emitted light as a function of magnetic field and temperature, it is found that the light polarization exactly corresponds to the magnetization of the (Ga,Mn)As layer. After performing the required control measurements, this clearly demonstrated spin injection from the (Ga,Mn)As ferromagnetic layer into the non-magnetic GaAs layer.

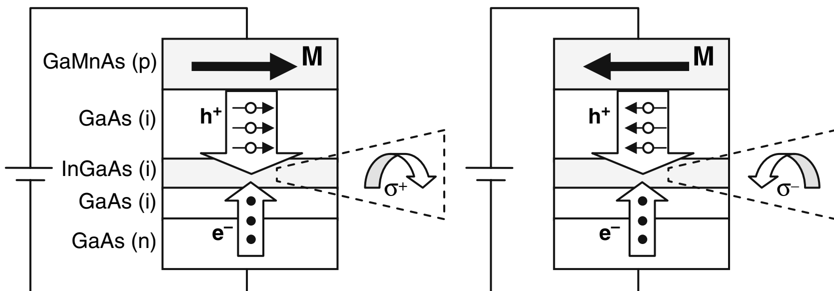


Fig. 5.29 Dependence of the light helicity on the spin injection in a spin-LED

Following the demonstration of electric spin injection in all-semiconductor structures, Schmidt et al. [138] provided an explanation for the success of these structures. In a model which assumes diffusive transport and Ohm's law (including ohmic contacts), the spin polarization of the injected carriers is found to be $\beta(\sigma_N/\sigma_F)(\lambda_F/\lambda_N)$, where β is the spin polarization of the ferromagnet, σ_F and σ_N are the conductivities of the ferromagnet and non-magnet, respectively, and λ_F and λ_N are the spin-diffusion lengths of the ferromagnet and non-magnet, respectively. The problem encountered for spin injection from a ferromagnetic *metal* into a non-magnetic *semiconductor* is that the conductance ratio σ_N/σ_F is very small, on the order of 0.001. Thus a ferromagnet with $\beta \sim 30\%$ will generate a spin polarization of only about 0.03% in the semiconductor. The success of the all-semiconductor structures is due to the fact that there is no serious conductivity mismatch. It is interesting to note that the conductivity mismatch term is present in earlier work related to all-metal structures, but was never emphasized because it was never a source of problems in those structures [63].

Soon afterward, one solution to the conductivity mismatch problem was provided by Rashba [139] and by Fert and Jaffres [140]. By introducing a tunnel barrier between the ferromagnetic metal and the non-magnetic semiconductor, high spin injection efficiency can be achieved if the barrier's resistance is larger than the semiconductor's. In some metal/semiconductor systems, there exists a potential barrier on the semiconductor side of the interface known as the Schottky barrier. Its height and width depend on many factors including the work functions of the metal and semiconductor, interface states, the bandgap of the semiconductor, and the doping type and concentration. In Fe/GaAs, the presence of a Schottky barrier makes spin injection from Fe into GaAs possible without introducing an oxide tunnel barrier, as was demonstrated through spin-LED experiments [141, 142, 143]. It was later shown that using doping gradients to adjust the Schottky barrier width could enhance the spin injection efficiency [144]. Finally, AlO_x and MgO tunnel barriers were introduced [145, 146], and the highest efficiency for spin injection was achieved in Fe/MgO/GaAs structures where the tunneling across the MgO leads to enhanced spin polarization due to the Δ_1 spin-filtering property discussed in Section 5.2.2.

5.3.3.2 Spin Extraction and Ferromagnetic Proximity Polarization

In a traditional spin valve such as CPP-GMR, spins are injected from a FM layer, transported across a NM layer, and detected by a second FM. In such devices, the spin polarization in the NM layer is generated through spin injection. An alternative method for generating spins in a NM layer is to utilize spin reflection or spin "extraction" [147]. Instead of *adding* spins to generate a spin polarization in the NM layer, the idea of extraction is to *remove* unwanted spins from the NM layer. A schematic drawing of this process is shown in Fig. 5.30a. The microscopic picture is based on the spin-dependent reflection at the

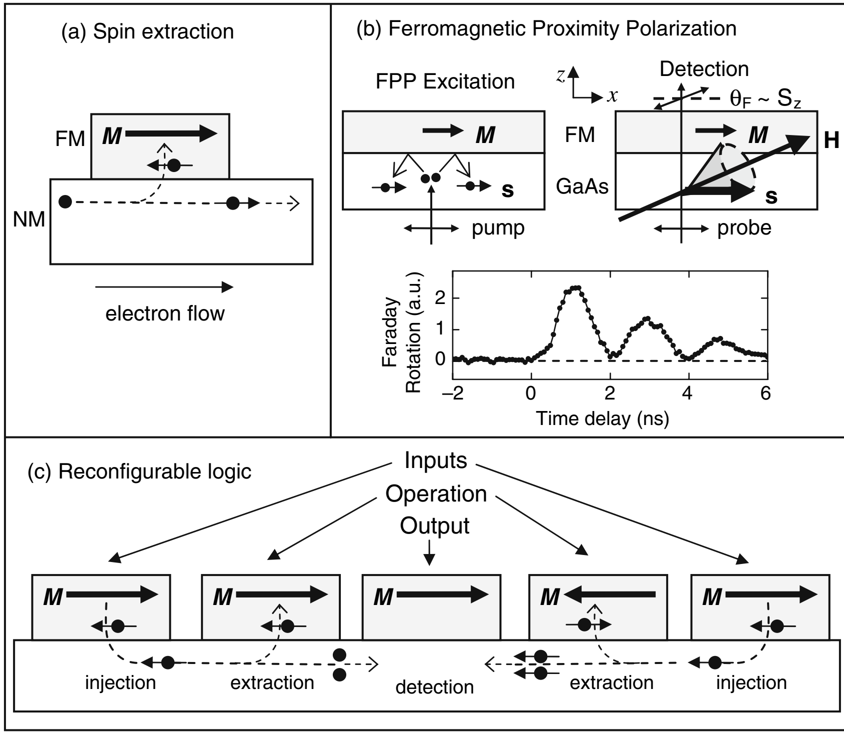


Fig. 5.30 (**a**) Spin extraction, (**b**) ferromagnetic proximity polarization, (**c**) proposed reconfigurable logic gate

NM/FM interface, which generates spin polarization in the NM. Such polarization is present in theoretical calculations of CPP-GMR [63] and contributes to the ultimate values of GMR in such systems. However, using this effect as a spin source was not explored at that time.

Experimentally, the direct measurement of spin polarization generated by reflection from a NM/FM interface was achieved through ultrafast optical measurements on MnAs/GaAs and Fe/GaAs systems [106, 107, 148]. Using a linearly polarized optical pump to generate unpolarized electrons in the GaAs layer, subsequent reflection of these electrons from the FM/GaAs interface generates a spin polarization in the GaAs layer that is parallel to the magnetization of the FM layer (Fig. 5.30b). These spins are detected by applying a magnetic field that lies slightly out of the GaAs plane and measuring the spin dynamics using TRFR. The FM magnetization remains in-plane due to the magnetic shape anisotropy, so the induced spin polarization in the GaAs is in-plane. Due to the angle between the spin and the applied field, the spins precess about the applied field in a cone shape. The time delay scan in Fig. 5.30b exhibits Faraday rotation which begins at zero (indicating that the spin is

in-plane), increases to a maxima (indicating that the spin has an out-of-plane component), returns to zero, and continues to oscillate. The presence of these oscillations is direct proof that the electrons in the GaAs are spin polarized, even though the photoexcitation is unpolarized. This optically driven process is known as ferromagnetic proximity polarization (FPP).

Subsequent optical experiments investigated lateral FM/GaAs devices under bias and directly measured the spin accumulation in the GaAs when the electrons flow from the GaAs into the FM [149, 150]. Under this “forward bias” condition, an unusual sign reversal of polarization as a function of bias was observed in FM/GaAs [38] and FM/Al_2O_3/Al [151].

Utilization of these phenomena in spintronic devices has been advocated in a number of device proposals [152, 153] and in a theory of spin extraction [147]. In particular, a reconfigurable logic circuit based on spin extraction and spin injection was recently proposed [153]. The reconfigurable logic gate consists of five ferromagnetic electrodes on top of a semiconducting channel (Fig. 5.30c). The outer two electrodes are the two inputs, the next two electrodes define the gate operation, and the center electrode is the output. Spins are injected from the two input electrodes and these spins flow to the center. The next two inner electrodes operate on the injected spins through spin extraction, and the resulting spin polarizations from the two branches add together at the center electrode, where it is read out. The logic operation can be reconfigured by changing the magnetization of the two inner FM electrodes either with magnetic field pulses or spin torque (like in MRAM). For an understanding of the logic operation, the reader is encouraged to read the original paper [153]. While it is unclear whether such a circuit will become a successful technology, this proposal illustrates the point that a computer based on spin can utilize physical principles that are inaccessible to purely charge-based electronics.

5.4 Lateral Spin Transport Devices

5.4.1 Lateral Spin Valves and Non-local Measurements

While much of the past technological successes of spintronics are related to the multilayered structures discussed in Section 5.2, there is significant interest in developing lateral spintronic devices. In a lateral geometry, multi-terminal devices are readily fabricated and the manipulation of spin via electrostatic gating becomes possible. The reconfigurable logic gate described in Fig. 5.30c and the Datta–Das spin transistor are some examples. Accompanying these potential advantages are new challenges which must be overcome. Primarily, spins must remain polarized for longer distances in lateral devices (over hundreds of nanometers) as compared to the multilayered

devices (a few nanometers). Therefore, alternative materials and more sensitive detection methods are desirable.

5.4.1.1 Lateral Spin Valve

The lateral spin valve (Fig. 5.31) is directly analogous to vertical spin valve devices such as the magnetic tunnel junction and the CPP-GMR. For both lateral and vertical orientations the current flows from one FM electrode to a second FM electrode. In lateral devices, spin-polarized electrons are injected from one FM electrode (injector), transported through a NM material, and flow into the second FM electrode (detector), as seen in Fig. 5.31. The characteristic signature of spin-polarized transport is the change in resistance between parallel and antiparallel magnetization alignments (i.e., magnetoresistance, MR). Modulations in the resistance of lateral devices are possible (in principle) via manipulation of spin by electrostatic gates in the Rashba effect [136], quantum interference effects [14, 15, 154], *g*-factor engineering [130], or other possible mechanisms.

Spin transport in carbon nanotube lateral spin valves exhibits an interesting dependence on gate voltage [14, 15, 154]. In one study, the MR has been found to oscillate between values of –7 and +17% as a function of gate voltage [14]. Such behavior is believed to originate from spin-dependent quantum interference effects caused by multiple reflections between the two FM contacts [154]. It is important to point out that the gate dependence of spin transport is a unique property for lateral devices which could not be realized in the multilayered devices discussed in Section 5.2.

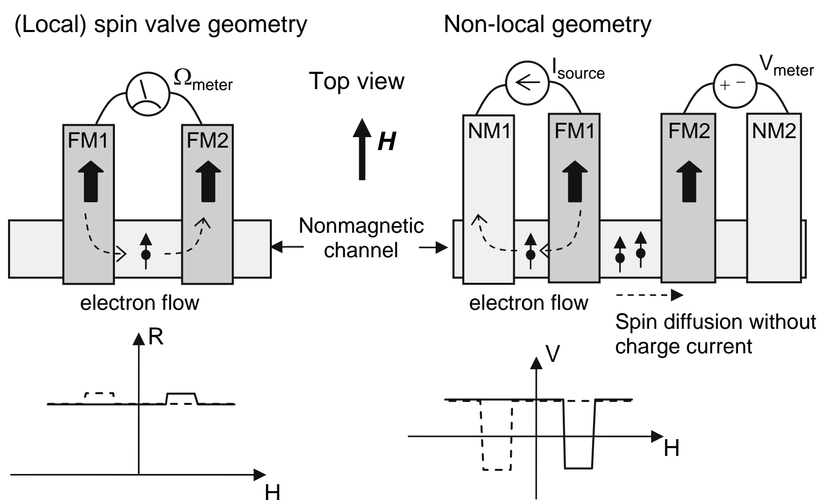


Fig. 5.31 Top view of lateral spin valves in a local and non-local measurement geometry, with schematic data for each geometry

5.4.1.2 Non-local Geometry

A more sensitive measurement of spin injection is the so-called "non-local" geometry, or the "Johnson–Silsbee" geometry [35, 155], which is commonly employed to identify spin injection and spin diffusion in lateral structures. Unlike the typical spin valve device, there is no current flow between the two FM electrodes. Instead, after the spins are injected from FM1 into the non-magnetic channel, the electrons are directed *away* from the FM2. However, through the phenomenon of spin diffusion, the spin polarization will spread and make its way to FM2 even though there is no electrical current between FM1 and FM2.

The idea of spin diffusion is not mysterious and is analogous to the diffusion of gas particles. Suppose you start with a box containing a gas of molecules A on the left and molecules B on the right, separated by a divider and having the same initial concentrations. When the divider is removed, the molecules move randomly and eventually the A and B species are uniformly distributed throughout the box. Even though there is no net particle flow, on average, A moves to the right and B moves to the left. This eventually leads to a uniform mixture of A and B. For the case of spin, consider a left region that has 100% spin-up polarization and a right region that is unpolarized. Due to standard electron diffusion, the left and right regions will exchange electrons so that there is no net electrical current. However, the electrons moving from left to right are 100% spin up, while the electrons moving from right to left are 50% spin up and 50% spin down. At the end of this process, the net effect is that spin has diffused to the right, even through there is no electrical current (Fig. 5.32).

In the non-local measurement, spins diffuse through the NM channel from FM1 to FM2. The spin polarization in the NM under the FM2 contact is detected by measuring the voltage between the NM and the FM2. A voltage is present because the spin polarization in the NM produces a spin-dependent chemical potential and the FM2 couples asymmetrically to this chemical potential. The net effect is that the voltage is positive or negative, depending on the relative orientation between the spin polarization in the NM and the magnetization of FM2 (Fig. 5.31). The non-local measurement was developed

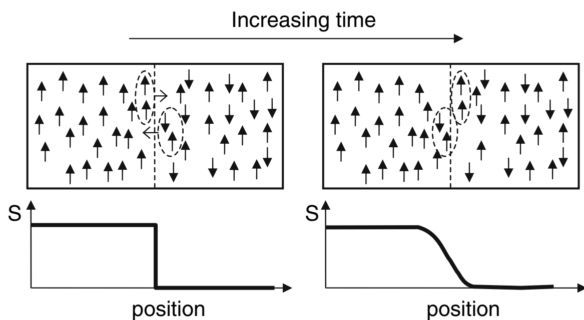


Fig. 5.32 A microscopic picture of spin diffusion with no overall charge current

in 1985 by Johnson and Silsbee to investigate spin injection into aluminum [35]. Many years later, with the advent of improved fabrication methods and new materials, the non-local measurements were performed in mesoscopic metal spin valves [36, 156, 157], semiconductors [38], carbon nanotubes [16], and graphene [19].

There are a couple of advantages of using the non-local detection method as compared to the conventional spin valve measurement. First, because this is not a resistance measurement, series resistance artifacts such as anisotropic magnetoresistance of the FM electrodes are eliminated. Second, non-local detection methods ideally exhibit no background level, greatly improving the signal-to-noise ratio and providing more sensitive spin detection.

5.4.1.3 Hanle Effect

The Hanle effect provides the clearest demonstration that the observed signals originate from spin injection. By applying an out-of-plane magnetic field ($H_\perp$), spin precession is induced in the injected electrons (Fig. 5.33a). In the non-local geometry, the spins reach FM2 through diffusion, so there is a large distribution of transit times, which leads to a distribution of spin orientations. The final spin polarization under FM2 depends on the spin precession, spin diffusion, and spin relaxation, and the dependence of this polarization is shown in the data and curve fits in Fig. 5.33b, taken from Ref. [36]. Qualitatively, the polarization is largest in zero field because all spins remain aligned. When $H_\perp$ is increased two effects occur. First, the spins precess at frequency $\omega_L = g\mu_B H_\perp/\hbar$ (see Equation 5.15), which promotes oscillations in the spin polarization at FM2 as a function of $H_\perp$. Second, because the transit times are broadly distributed for diffusion, the oscillatory behavior is washed out.

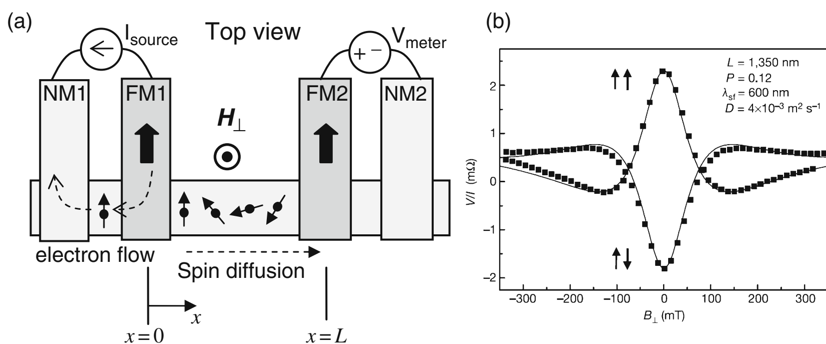


Fig. 5.33 (**a**) The Hanle effect: electrical detection of spin diffusion with precession. (**b**) Hanle data in all-metal lateral devices. Reprinted with permission from Macmillan Publishers Ltd: *Nature*, Ref. [36], copyright 2002

To understand the Hanle effect quantitatively, we need a quantitative understanding of diffusion. Diffusion is based on the random motion of particles. Mathematically, this is described by the diffusion equation:

$$\frac{\partial \rho}{\partial t} = D\nabla^2 \rho(\boldsymbol{r}, t) \tag{5.16}$$

where $\rho(\boldsymbol{r},t)$ is the density of the substance that is diffusing and D is the diffusion coefficient. In a one-dimensional problem, suppose a total of N particles are concentrated at the origin at $t = 0$: $\rho(x,t = 0) = N\delta(x)$, where $\delta(x)$ is the delta function. The solution for later times is given by

$$\rho(x, t) = \frac{N}{\sqrt{4\pi Dt}} \exp\left(-\frac{x^2}{4Dt}\right) \tag{5.17}$$

This solution is shown graphically in Fig. 5.34, where we take $D = 1$ and treat all quantities as dimensionless for clarity. In Fig. 5.34a, the density $\rho(x,t)$ spreads in position as time increases, as is expected intuitively for diffusion. In Fig. 5.34b, we consider the density at a fixed position ($x = 3$) as a function of time t. This curve represents the distribution of transit times for a particle starting at the origin and ending up at $x = 3$ at time t. The key point is that there is a broad distribution of transit times.

For the case of spin density, everything is the same except that there is spin relaxation due to spin flips. Unlike particle number, there is no conservation law for spin density. The diffusion equation for spin density (ρ_S) is

$$\frac{\partial \rho_S}{\partial t} = D\nabla^2 \rho_S(\boldsymbol{r}, t) - \frac{\rho_S}{\tau} \tag{5.18}$$

where the last term is due to spin flip scattering and τ is the characteristic time for spin flip (i.e., spin lifetime). Alternatively, one could write the diffusion

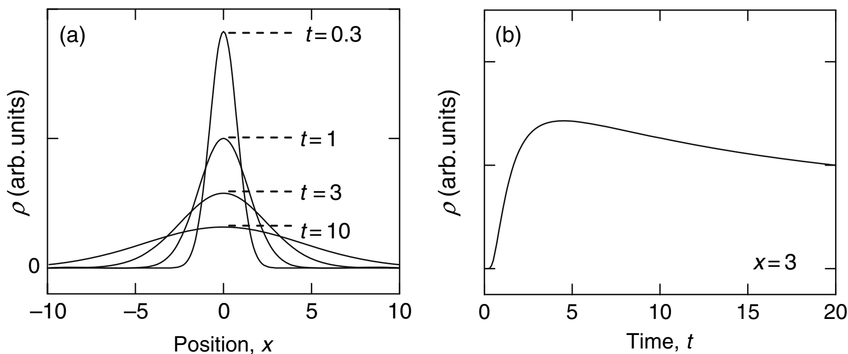


Fig. 5.34 (**a**) Solution to the diffusion equation at different times. The diffusion constant D is set to 1 for clarity. (**b**) The distribution of transit times of a particle starting at the origin and arriving at $x = 3$

equation in terms of spin-dependent chemical potentials, where $\rho_S \sim \Delta\mu = \mu_\uparrow - \mu_\downarrow$, but we keep with our current notation. In a one-dimensional problem, suppose that the spin density is concentrated at the origin at $t = 0$: $\rho_S(x,t = 0) = \delta(x)$, where $\delta(x)$ is the delta function. The solution for later times is

$$\rho_S(x,t) = \frac{1}{\sqrt{4\pi Dt}} \exp\left(-\frac{x^2}{4Dt}\right) \exp\left(-\frac{t}{\tau}\right) \tag{5.19}$$

This relation takes into account both the spin diffusion and the spin relaxation. The last remaining ingredient for the Hanle effect is the spin precession. For concreteness, let us assume that FM1 ($x = 0$) injects a spin-up electron, which precesses at a frequency of $\omega_L = g\mu_B H_\perp/\hbar$. The component of spin along the original axis (up) is given by $\cos(g\mu_B H_\perp t/\hbar)$. Clearly, the contribution of this electron to the overall spin polarization beneath FM2 ($x = L$) depends on its transit time. Thus, to calculate the total spin polarization beneath FM2, we need to sum over all contributions for all transit times from FM1 ($x = 0$) to FM2 ($x = L$). This is given by

$$\begin{aligned} \mathrm{SP}(x = L) &\sim \int_0^\infty \rho_S(L,t) \cos(g\mu_B H_\perp t/\hbar) \mathrm{d}t \\ &= \int_0^\infty \frac{1}{\sqrt{4\pi Dt}} \exp\left(-\frac{L^2}{4Dt}\right) \exp\left(-\frac{t}{\tau}\right) \cos(g\mu_B H_\perp t/\hbar) \mathrm{d}t \end{aligned} \tag{5.20}$$

The three factors for spin diffusion, relaxation, and precession are evident in the final expression. This equation is used to fit the Hanle data and is the solid line in Fig. 5.33b.

The Hanle curve in Fig. 5.33b has a strong peak at zero field, but at higher $H_\perp$ the oscillations due to spin precession are washed out due to the wide range of transit times associated with spin diffusion (Fig. 5.34). To achieve oscillatory electrical signals, such as those needed for a Datta–Das spin transistor (Fig. 5.27), the spins should be driven by an electrical bias (e.g., local spin valve geometry) so that the transit time between FM1 and FM2 is more uniform. In this mode, the spins are transported by electron drift. For spintronic device applications, spin precession under drift conditions can produce the desired oscillatory signals. Such behavior has recently been observed in spin precession in silicon under drift conditions [158].

5.4.2 Spin Hall Effect

The spin Hall effect is a manifestation of spin–orbit coupling and has recently been observed in both semiconductor and metallic systems [159, 160, 161, 162,163]. The basic behavior of the spin Hall effect is shown in Fig. 5.35.

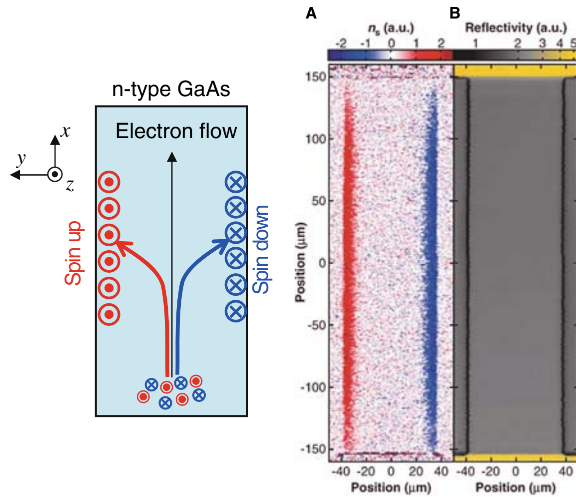


Fig. 5.35 (*Left*) General behavior of the spin Hall effect, where an unpolarized charge current generates a transverse spin current. (*Right*) Magneto-optical imaging of the spin Hall effect in n-type GaAs, from Ref. [159]. Reprinted with permission from AAAS (*See* Color Insert)

A charge current along the x-direction generates a pure spin current along the transverse direction (y-axis), leading to the accumulation of spin-up and spin-down electrons (with respect to the z-axis) at the edges of the film. A "pure spin current" implies that spin up is moving to the left while an equal amount of spin down is moving to the right, resulting in a spin current without a charge current. In both metals and semiconductors, such behavior can be generated by scattering from impurities with high atomic numbers (due to their strong spin–orbit coupling). When the spin Hall effect is generated by scattering from impurities, it is said to be "extrinsic" [164, 165, 166, 167]. For the case of p-type semiconductors, the spin Hall effect could be generated by the spin–orbit coupling present in the ideal band structure of the valence band [168, 169]. This effect is said to be "intrinsic" due to the fact that it is not related to the presence of defects or impurities. In our discussion, we will focus on the extrinsic spin Hall effect because it could be observed in many systems and its strength should be adjustable by controlled doping. In addition, in this qualitative discussion, we will not be careful about the overall sign of the spin Hall effect (for example, it depends on the g-factor which has opposite signs for GaAs and Al).

The spin Hall effect was first discovered in semiconductor systems using Kerr microscopy on a strip of n-type GaAs (Fig. 5.35) [159]. In a second experiment, the spin accumulation at the edges of a semiconductor was detected by analyzing the circular polarization of electroluminescence intensity (i.e., spin-LED method) [160]. It was shown that at opposite edges, the circular polarization is of opposite sign and also depends on the sign of the current. This behavior is characteristic of the spin Hall effect.

In metallic systems where the optical detection methods are less sensitive, the spin Hall effect was detected by electrical measurements [161, 162, 163]. If one begins with a spin-polarized current along the x-axis, the asymmetric deflection

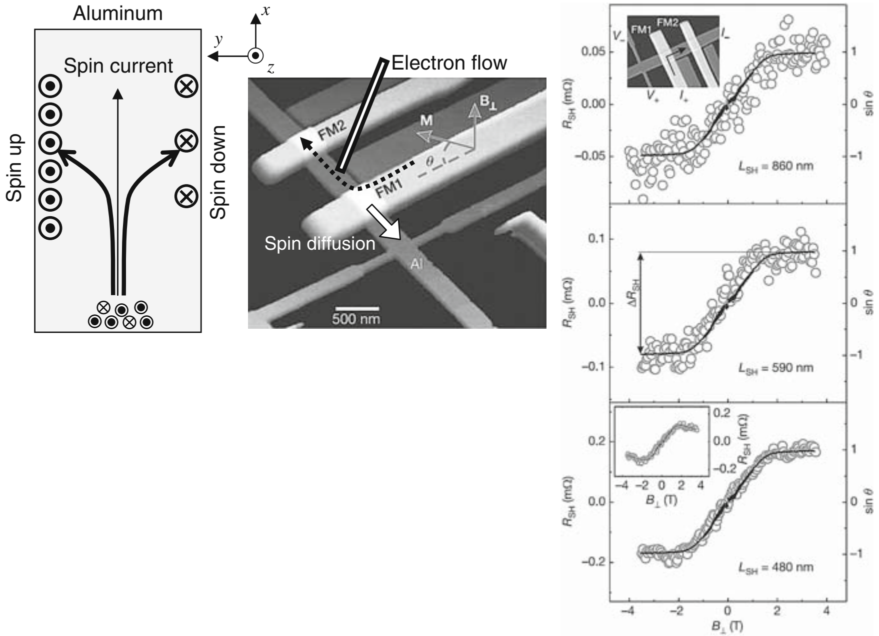


Fig. 5.36 (*Left*) General behavior of the inverse spin Hall effect, where a spin current generates a transverse voltage. (*Center and Right*) Experimental geometry and measurement of the inverse spin Hall effect in aluminum. Reprinted with permission from Macmillan Publishers Ltd: *Nature*, Ref. [161], copyright 2006

of spin-up and spin-down electrons will generate a lateral voltage, as shown in Fig. 5.36. This effect has been called the "inverse spin Hall effect" because a spin current generates a transverse voltage, while in the normal spin Hall effect a charge current generates a transverse spin accumulation. However, both effects have the same origin in asymmetric spin scattering from heavy impurities. The measurement in Fig. 5.36 is rather clever in that there is no charge current between the injection point and the lateral voltage electrodes. The current is actually directed away from the lateral voltage electrodes, causing the spin polarization to move toward the lateral electrode only by spin diffusion (like a non-local measurement). This is important because a charge current would generate a normal Hall voltage, which would complicate the analysis. As the magnetic field is ramped along the out-of-plane direction, the lateral voltage is found to trace the out-of-plane magnetization of the FM spin injector, which is the direct measurement of inverse spin Hall effect.

The explanation of the extrinsic spin Hall effect can be seen by considering the scattering from an impurity (Fig. 5.37). If the electron passes to the right of the impurity, it will be deflected toward the left due to the attractive interaction between the electron and the impurity atom. During this "fly-by," the potential energy gradient (represented by the contour lines) results in an internal

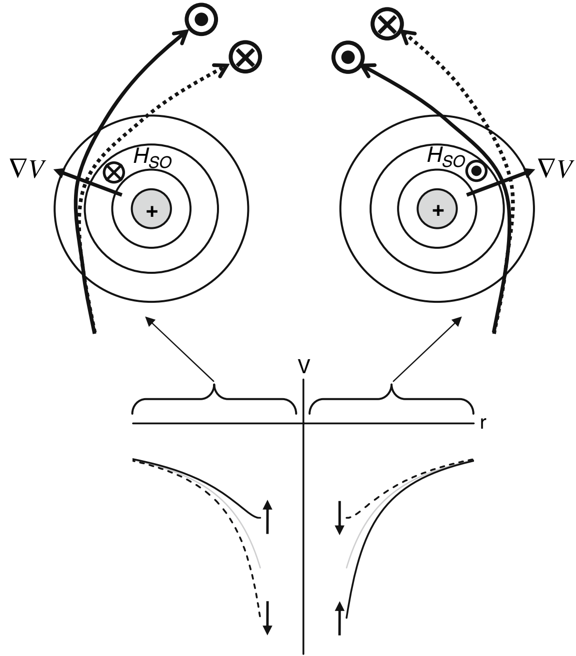


Fig. 5.37 Spin-orbit mechanism for extrinsic spin Hall effect due to impurity scattering. (*Top*) Spin-dependent trajectories of electrons around impurities. (*Bottom*) Spin-dependent potential energy for spin up (*solid black line*), spin down (*dotted line*). The *gray curve* is the Coulomb potential without the effect of spin-orbit coupling

spin–orbit magnetic field $\boldsymbol{H}_{\mathrm{SO}} \sim \nabla V \times \boldsymbol{k}$, which points out of the plane of the paper. $\boldsymbol{H}_{\mathrm{SO}}$ generates a spin-splitting between spin-up and spin-down states. The difference in the energies is given by $\Delta E = E_{\uparrow}–E_{\downarrow} = g\mu_{\mathrm{B}}\boldsymbol{H}_{\mathrm{SO}}$. The total scattering potential including spin–orbit effects is shown at the bottom of Fig. 5.37. The scattering potential and the final trajectories of the two spin states will depend on the sign of the spin and the side on which they fly by the impurity. In the following analysis, we assume that $g < 0$ (as in GaAs). When the electron passes to the right of the impurity we find $\Delta E < 0$, or equivalently $E_{\uparrow} < E_{\downarrow}$. This means that the spin-up electron experiences a deeper scattering potential than the spin-down electron (Fig. 5.37, bottom), so that there is a greater deflection for a spin-up electron than the spin-down electron (Fig. 5.37, top). If instead the electron passes to the left of the impurity, the gradient ∇V is in the opposite direction of the previous case, so that the $\boldsymbol{H}_{\mathrm{SO}}$ is also of opposite sign, resulting in a $\Delta E > 0$, or $E_{\downarrow} < E_{\uparrow}$. In this case, the spin-down electron experiences a deeper scattering potential and thus makes a sharper turn. In Fig. 5.37 (bottom) the potential energy of spin up is indicated by solid black lines and spin down is indicated by dashed lines, with the Coulomb potential energy in gray. The right half of the figure corresponds to when the electron flies by to the right of the impurity, while the left half represents when the electron flies by to the left. In either case, the spin-up electron goes more to the left, while the spin-down electron goes more to the right. Thus, in a sample such as the one shown in Fig. 5.35 where there is an average electron flow in the $+x$-direction, the difference in the

trajectories due to $\boldsymbol{H}_{\mathrm{SO}}$ causes spin-up electrons to be deflected more toward the left compared to spin-down, and spin-down electrons are deflected more to the right compared to spin-up. This results in a net spin current in the transverse direction without any associated charge current. The spin Hall effect represents a method of using the spin–orbit coupling to generate pure spin currents and spin polarization without the use of ferromagnets.

5.5 Concluding Remarks

We conclude by mentioning that this tutorial is not an exhaustive review of the field of spintronics and many important topics were not covered. For example, the leading candidate for quantum computing using spins—spins on localized states such as quantum dots or point defects—was not discussed. We have focused on the topics of spin transport, spin dynamics, and the roles of ferromagnets and spin–orbit coupling.

Acknowledgments RKK thanks his former research advisors—Z. Q. Qiu, D. D. Awschalom, and A. C. Gossard—for providing him with the opportunity to pursue research in these areas.

5.6 Appendix: Quantum Mechanics of Spin-1/2

We will review some key properties of quantum spin operators for the case of spin-1/2 particles. Operators are denoted by a tilde and real-space vectors are denoted by boldface lettering. The spin operators $\tilde{S}_x, \tilde{S}_y, \tilde{S}_z$ obey the commutation relations:

$$[\tilde{S}_x, \tilde{S}_y] = \mathrm{i}\hbar\tilde{S}_z,\ [\tilde{S}_y, \tilde{S}_z] = \mathrm{i}\hbar\tilde{S}_x,\ [\tilde{S}_z, \tilde{S}_x] = \mathrm{i}\hbar\tilde{S}_y \tag{5.21}$$

where the commutator is defined in general as $[\tilde{A}, \tilde{B}] \equiv \tilde{A}\tilde{B} - \tilde{B}\tilde{A}$.

While all the properties can be derived from these relations, we will only cite the key results that will be useful and refer readers interested in the detailed derivations to textbooks.

The spin magnitude operator is given by $\tilde{\boldsymbol{S}}^2 = \tilde{S}_x^2 + \tilde{S}_y^2 + \tilde{S}_z^2$. Using relations (5.21), one can derive that $[\tilde{\boldsymbol{S}}^2, \tilde{S}_x] = [\tilde{\boldsymbol{S}}^2, \tilde{S}_y] = [\tilde{\boldsymbol{S}}^2, \tilde{S}_z] = 0$, which implies that quantum numbers for $\tilde{\boldsymbol{S}}^2$ and $\tilde{\boldsymbol{S}}$ projected along one axis can be defined simultaneously (the projected axis is traditionally chosen as the z-axis). Thus, the quantum spin states are $|s, m_s\rangle$, where s is the spin magnitude quantum number and m_s is the z-component quantum number, and the eigenvalue relations are

$$\tilde{\boldsymbol{S}}^2|s, m_s\rangle = \hbar^2 s(s+1)|s, m_s\rangle \tilde{S}_z|s, m_s\rangle = \hbar m_s|s, m_s\rangle \tag{5.22}$$

For spin angular momentum, s can take on positive half-integer values, and m_s can take on half-integer values between $-s$ and s. For electrons, $s = ½$ and

m_s can be $+\frac{1}{2}$ or $-\frac{1}{2}$. A conventional notation is $|\uparrow\rangle \equiv |s = \frac{1}{2}, m_s = \frac{1}{2}\rangle$ and $|\downarrow\rangle \equiv |s = \frac{1}{2}, m_s = -\frac{1}{2}\rangle$.

For performing calculations, it is useful to define raising ($\tilde{S}_+$) and lowering ($\tilde{S}_-$) operators: $\tilde{S}_+ = \tilde{S}_x + \mathrm{i}\tilde{S}_y$ and $\tilde{S}_- = \tilde{S}_x - \mathrm{i}\tilde{S}_y$. The inverse relations are $\tilde{S}_x = \frac{\tilde{S}_+ + \tilde{S}_-}{2}$ and $\tilde{S}_y = \frac{\tilde{S}_+ - \tilde{S}_-}{2\mathrm{i}}$. The raising and lowering operators have the property that they increase or decrease the m_s quantum number by one, according to the relations:

$$\tilde{S}_\pm|s, m_s\rangle = \hbar\sqrt{s(s+1) - m_s(m_s \pm 1)}|s, m_s \pm 1\rangle \tag{5.23}$$

For electrons, this is summarized by

$$\tilde{S}_+|\downarrow\rangle = \hbar|\uparrow\rangle,\ \tilde{S}_+|\uparrow\rangle = 0,\ \tilde{S}_-|\uparrow\rangle = \hbar|\downarrow\rangle,\ \tilde{S}_-|\downarrow\rangle = 0 \tag{5.24}$$

In order to develop a classical interpretation of the spin states, it is useful to calculate expectation values of the spin operator $\tilde{\boldsymbol{S}}$. Let us consider the $|\uparrow\rangle$ state:

$$\begin{aligned}
\langle\uparrow|\tilde{S}_z|\uparrow\rangle &= \left\langle\uparrow\left|\frac{\hbar}{2}\right|\uparrow\right\rangle = \frac{\hbar}{2} \\
\langle\uparrow|\tilde{S}_x|\uparrow\rangle &= \left\langle\uparrow\left|\frac{\tilde{S}_+ + \tilde{S}_-}{2}\right|\uparrow\right\rangle = \left\langle\uparrow\left|\frac{\tilde{S}_-}{2}\right|\uparrow\right\rangle = \frac{\hbar}{2}\langle\uparrow|\downarrow\rangle = 0 \\
\langle\uparrow|\tilde{S}_y|\uparrow\rangle &= \left\langle\uparrow\left|\frac{\tilde{S}_+ - \tilde{S}_-}{2\mathrm{i}}\right|\uparrow\right\rangle = \left\langle\uparrow\left|\frac{-\tilde{S}_-}{2\mathrm{i}}\right|\uparrow\right\rangle = -\frac{\hbar}{2\mathrm{i}}\langle\uparrow|\downarrow\rangle = 0
\end{aligned} \tag{5.25}$$

Written in terms of unit vectors $\hat{i}, \hat{j}, \hat{k}$ along the x,y,z-axes, the expectation value is $\langle\uparrow|\tilde{S}_x\hat{i} + \tilde{S}_y\hat{j} + \tilde{S}_z\hat{k}|\uparrow\rangle = \frac{\hbar}{2}\hat{k}$. Thus, the $|\uparrow\rangle$ state can be interpreted classically as a spin vector along the z-axis.

A more complicated situation is to consider a spin oriented along an arbitrary direction. Consider an arbitrary axis labeled z', which has polar and azimuthal angles of θ and ϕ with respect to the x,y,z-axes, as shown in Fig. 5.38. We

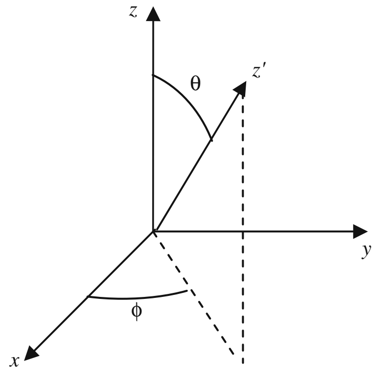


Fig. 5.38 Polar and azimuthal angles

propose without justification that the spin-up eigenstate along the z' axis is given by

$$|\uparrow\rangle_{z'} = \cos(\theta/2)\exp(-i\phi/2)|\uparrow\rangle_z + \sin(\theta/2)\exp(i\phi/2)|\downarrow\rangle_z \tag{5.26}$$

where the subscripts z and z' identify the quantization axis. To be convinced that this represents a spin along the z' axis, we just calculate the expectation value of $\tilde{S}$:

$$
\begin{aligned}
{}_{z'}\langle\uparrow|\tilde{S}_z|\uparrow\rangle_{z'} &= \frac{\hbar}{2}\cos^2(\theta/2) - \frac{\hbar}{2}\sin^2(\theta/2) = \frac{\hbar}{2}\cos(\theta) \\
{}_{z'}\langle\uparrow|\tilde{S}_x|\uparrow\rangle_{z'} &= {}_{z'}\left\langle\uparrow\left|\frac{\tilde{S}_+ + \tilde{S}_-}{2}\right|\uparrow\right\rangle_{z'} = \frac{\hbar}{2}\cos(\theta/2)\sin(\theta/2)\exp(i\phi) \\
&\quad + \frac{\hbar}{2}\cos(\theta/2)\sin(\theta/2)\exp(-i\phi) = \frac{\hbar}{2}\sin(\theta)\cos(\phi) \\
{}_{z'}\langle\uparrow|\tilde{S}_y|\uparrow\rangle_{z'} &= {}_{z'}\left\langle\uparrow\left|\frac{\tilde{S}_+ - \tilde{S}_-}{2i}\right|\uparrow\right\rangle_{z'} = \frac{\hbar}{2i}\cos(\theta/2)\sin(\theta/2)\exp(i\phi) \\
&\quad - \frac{\hbar}{2i}\cos(\theta/2)\sin(\theta/2)\exp(-i\phi) = \frac{\hbar}{2}\sin(\theta)\sin(\phi)
\end{aligned}
$$

Thus, the expectation value of the spin operator is

$${}_{z'}\langle\uparrow|\tilde{S}_x\hat{i} + \tilde{S}_y\hat{j} + \tilde{S}_z\hat{k}|\uparrow\rangle_{z'} = \frac{\hbar}{2}(\sin(\theta)\cos(\phi)\hat{i} + \sin(\theta)\sin(\phi)\hat{j} + \cos(\theta)\hat{k}) \tag{5.27}$$

which is a vector oriented along the z' axis. Thus, the state in Equation 5.26 is a spin vector along the z' direction.

Finally, note that for the special case of z' being the x-axis ($\theta = 90°$, $\phi = 0°$), we get

$$|\uparrow\rangle_x = \frac{|\uparrow\rangle_z + |\downarrow\rangle_z}{\sqrt{2}} \tag{5.28}$$

References

1. G. A. Prinz, Science **282**, 1660 (1998).
2. S. A. Wolf, D. D. Awschalom, R. A. Buhrman, et al., Science **294**, 1488 (2001).
3. D. D. Awschalom, N. Samarth, and D. Loss, *Semiconductor Spintronics and Quantum Computation* (Springler-Verlag, Berlin, 2002).
4. J. M. D. Coey, M. Viret, and S. Von Molnar, Adv. Phys. **48**, 167 (1999).
5. M. B. Salamon and M. Jaime, Rev. Mod. Phys. **73**, 583 (2001).
6. V. Dediu, M. Murgia, E. C. Matacotta, et al., Solid State Commun. **122**, 181 (2002).
7. Z. H. Xiong, D. Wu, Z. V. Vardeny, et al., Nature **427**, 821 (2004).

8. J. Petta, S. K. Slater, and D. C. Ralph, Phys. Rev. Lett. **93**, 136601 (2004).
9. S. Majumdar, H. S. Majumdar, R. Laiho, et al., J. Alloys Compd. **423**, 169 (2006).
10. S. Pramanik, C.-G. Stefanita, S. Patibandla, et al., Nat. Nanotechnol. **2**, 216 (2007).
11. K. Tsukagoshi, B. W. Alphenaar, and H. Ago, Nature **401**, 572 (1999).
12. B. Zhao, I. Monch, H. Vinzelberg, et al., Appl. Phys. Lett. **80**, 3144 (2002).
13. J.-R. Kim, H. M. So, J.-J. Kim, et al., Phys. Rev. B **66**, 233401 (2002).
14. S. Sahoo, T. Kontos, J. Furer, et al., Nat. Phys. **1**, 99 (2005).
15. H. T. Man, J. W. Wever, and A. F. Morpurgo, Phys. Rev. B (Rapid Commun.) **73**, 241401 (2006).
16. N. Tombros, S. J. van der Molen, and B. J. van Wees, Phys. Rev. B **73**, 233403 (2006).
17. R. Thamankar, S. Niyogi, B. Y. Yoo, et al., Appl. Phys. Lett. **89**, 033119 (2006).
18. L. E. Hueso, J. M. Pruneda, V. Ferrari, et al., Nature **445**, 410 (2007).
19. N. Tombros, C. Jozsa, M. Popinciuc, et al., Nature **448**, 571 (2007).
20. M. Nishioka and A. M. Goldman, Appl. Phys. Lett. **90**, 252505 (2007).
21. W. H. Wang, K. Pi, Y. Li, et al., Phys. Rev. B (Rapid Commun.) **77**, 020402 (2008).
22. S. Cho, Y.-F. Chen, and M. S. Fuhrer, Appl. Phys. Lett. **91**, 123105 (2007).
23. M. N. Baibich, J. M. Broto, A. Fert, et al., Phys. Rev. Lett. **61**, 2472 (1988).
24. J. S. Moodera, L. R. Kinder, T. M. Wong, et al., Phys. Rev. Lett. **74**, 3273 (1995).
25. E. B. Myers, D. C. Ralph, J. A. Katine, et al., Science **285**, 867 (1999).
26. P. A. M. Dirac, *The Principle Of Quantum Mechanics* (Oxford University Press, Oxford, UK, 1958).
27. P. W. Shor, *Proceedings of the 35th Annual Symposium on the Foundations of Computer Science* (IEEE Computer Society Press, Los Alamitos, CA, 1994), p. 124.
28. L. K. Grover, Phys. Rev. Lett. **79**, 325 (1997).
29. C. H. Bennett and D. P. DiVincenzo, Nature **404**, 247 (2000).
30. L. M. K. Vandersypen, S. M., G. Breyta, et al., Nature **414**, 883 (2001).
31. J. M. Kikkawa and D. D. Awschalom, Phys. Rev. Lett. **80**, 4313 (1998).
32. J. M. Kikkawa, I. P. Smorchkova, N. Samarth, et al., Science **277**, 1284 (1997).
33. S. Ghosh, V. Sih, W. H. Lau, et al., Appl. Phys. Lett. **86**, 232507 (2005).
34. J. Z. Sun, Phys. Rev. B **62**, 570 (2000).
35. M. Johnson and R. H. Silsbee, Phys. Rev. Lett. **55**, 1790 (1985).
36. F. J. Jedema, H. B. Heersche, A. T. Filip, et al., Nature **416**, 713 (2002).
37. P. R. Hammar and M. Johnson, Phys. Rev. Lett. **88**, 066806 (2002).
38. X. Lou, C. Adelmann, S. A. Crooker, et al., Nat. Phys. **3**, 197 (2007).
39. P. Grunberg, R. Schreiber, Y. Pang, et al., Phys. Rev. Lett. **57**, 2442 (1986).
40. P. M. Tedrow and R. Meservey, Phys. Rev. Lett. **26**, 192 (1971).
41. M. Julliere, Phys. Lett. **54A**, 225 (1974).
42. S. S. P. Parkin, IBM J. Res. Dev. **42**, 3 (1998).
43. M. A. Ruderman and C. Kittel, Phys. Rev. **96**, 99 (1954).
44. T. Kasuya, Prog. Theor. Phys. **16**, 45 (1956).
45. K. Yosida, Phys. Rev. **106**, 893 (1957).
46. S. S. P. Parkin, Phys. Rev. Lett. **67**, 3598 (1991).
47. P. Bruno and C. Chappert, Phys. Rev. Lett. **67**, 1602 (1991).
48. P. Bruno and C. Chappert, Phys. Rev. B **46**, 261 (1992).
49. J. Unguris, R. J. Celotta, and D. T. Pierce, Phys. Rev. Lett. **67**, 140 (1991).
50. Z. Q. Qiu, J. Pearson, A. Berger, et al., Phys. Rev. Lett. **68**, 1398 (1992).
51. P. J. H. Bloemen, M. T. Johnson, M. T. H. van de Vorst, et al., Phys. Rev. Lett. **72**, 764 (1994).
52. M. D. Stiles, Phys. Rev. B **48**, 7238 (1993).
53. P. Bruno, Phys. Rev. B **52**, 411 (1995).
54. J. E. Ortega and F. J. Himpsel, Phys. Rev. Lett. **69**, 844 (1992).
55. R. K. Kawakami, E. Rotenberg, E. J. Escorcia-Aparicio, et al., Phys. Rev. Lett. **80**, 1754 (1998).

56. R. E. Camley and J. Barnas, Phys. Rev. Lett. **63**, 664 (1989).
57. P. M. Levy, S. Zhang, and A. Fert, Phys. Rev. Lett. **65**, 1643 (1990).
58. D. M. Edwards, J. Mathon, R. B. Muniz, et al., Phys. Rev. Lett. **67**, 493 (1991).
59. J. M. George, L. G. Pereira, A. Barthelemy, et al., Phys. Rev. Lett. **72**, 408 (1994).
60. J. P. Renard, P. Bruno, R. Megy, et al., Phys. Rev. B. **51**, 12821 (1995).
61. S. S. P. Parkin, Phys. Rev. Lett. **71**, 1641 (1993).
62. W. P. Pratt Jr., S.-F. Lee, J. M. Slaughter, et al., Phys. Rev. Lett. **66**, 3060 (1991).
63. T. Valet and A. Fert, Phys. Rev. B **48**, 7099 (1993).
64. T. Miyazaki and N. Tezuka, J. Magn. Magn. Mater. **139**, L231 (1995).
65. M. Bowen, V. Cros, F. Petroff, et al., Appl. Phys. Lett. **79**, 1655 (2001).
66. J. Faure-Vincent, C. Tiusan, E. Jouguelet, et al., Appl. Phys. Lett. **82**, 4507 (2003).
67. S. S. P. Parkin, C. Kaiser, A. Panchula, et al., Nat. Mater. **3**, 862 (2004).
68. S. Yuasa, T. Nagahama, A. Fukushima, et al., Nat. Mater. **3**, 868 (2004).
69. S. Yuasa, A. Fukushima, H. Kubota, et al., Appl. Phys. Lett. **89**, 042505 (2006).
70. P. Mavropoulos, N. Papanikolaou, and P. H. Dederichs, Phys. Rev. Lett. **85**, 1088 (2000).
71. W. H. Butler, X.-G. Zhang, T. C. Schulthess, et al., Phys. Rev. B **63**, 054416 (2001).
72. J. Mathon and A. Umerski, Phys. Rev. B (Rapid Commun.) **63**, 220403 (2001).
73. X.-G. Zhang and W. H. Butler, Phys. Rev. B **70**, 172407 (2004).
74. L. Berger, Phys. Rev. B **54**, 9353 (1996).
75. J. C. Slonczewski, J. Magn. Magn. Mater. **159**, L1 (1996).
76. M. Tsoi, A. G. M. Jansen, J. Bass, et al., Phys. Rev. Lett. **80**, 4281 (1998).
77. M. D. Stiles and A. Zangwill, Phys. Rev. B **66**, 014407 (2002).
78. J. A. Katine, F. J. Albert, R. A. Buhrman, et al., Phys. Rev. Lett. **84**, 3149 (2000).
79. I. Krivorotov, N. C. Emley, J. C. Sankey, et al., Science **307**, 228 (2005).
80. Y. Acremann, J. P. Strachan, V. Chembrolu, et al., Phys. Rev. Lett. **96**, 217202 (2006).
81. S. Kaka, M. R. Pufall, W. H. Rippard, et al., Nature **437**, 389 (2005).
82. F. B. Mancoff, N. D. Rizzo, B. N. Engel, et al., Nature **437**, 393 (2005).
83. Z. Diao, D. Apalkov, M. Pakala, et al., Appl. Phys. Lett. **87**, 232502 (2005).
84. G. D. Fuchs, J. A. Katine, S. I. Kiselev, et al., Phys. Rev. Lett. **96**, 186603 (2006).
85. N. Vernier, D. A. Allwood, D. Atkinson, et al., Europhys. Lett. **65**, 526 (2004).
86. A. Yamaguchi, T. Ono, S. Nasu, et al., Phys. Rev. Lett. **92**, 077205 (2004).
87. M. Yamanouchi, D. Chiba, F. Matsukura, et al., Nature **428**, 539 (2004).
88. E. Saitoh, H. Miyajima, T. Yamaoka, et al., Nature **432**, 203 (2004).
89. M. Klaui, P.-O. Jubert, R. Allenspach, et al., Phys. Rev. Lett. **95**, 026601 (2005).
90. L. Thomas, M. Hayashi, X. Jiang, et al., Science **315**, 1553 (2007).
91. T. Kimura, Y. Otani, and J. Hamrle, Phys. Rev. Lett. **96**, 037201 (2006).
92. W. H. Meiklejohn and C. P. Bean, Phys. Rev. **102**, 1413 (1956).
93. W. H. Meiklejohn and C. P. Bean, Phys. Rev. **105**, 904 (1957).
94. J. Nogues and I. K. Schuller, J. Magn. Magn. Mater. **192**, 203 (1999).
95. A. E. Berkowitz and K. Takano, J. Magn. Magn. Mater. **200**, 552 (1999).
96. W. J. Gallagher, S. S. Parkin, Y. Lu, et al., J. Appl. Phys. **81**, 3741 (1997).
97. B. N. Engel, J. Akerman, B. Butcher, et al., IEEE Trans. Magn. **41**, 132 (2005).
98. J. Akerman, Science **308**, 508 (2005).
99. W. J. Gallagher and S. S. P. Parkin, IBM J. Res. Dev. **50**, 5 (2006).
100. H. Ohno, Science **281**, 951 (1998).
101. H. Ohno, D. Chiba, F. Matsukura, et al., Nature **408**, 944 (2000).
102. H. Boukari, P. Kossacki, M. Bertolini, et al., Phys. Rev. Lett. **88**, 207204 (2002).
103. S. Koshihara, A. Oiwa, M. Hirasawa, et al., Phys. Rev. Lett. **78**, 4617 (1997).
104. R. Fiederling, M. Keim, G. Reuscher, et al., Nature **402**, 787 (1999).
105. Y. Ohno, D. K. Young, B. Beschoten, et al., Nature **402**, 790 (1999).
106. R. K. Kawakami, Y. Kato, M. Hanson, et al., Science **294**, 131 (2001).
107. R. J. Epstein, I. Malajovich, R. K. Kawakami, et al., Phys. Rev. B **65**, 121202 (2002).

108. H. Ohno, H. Munekata, T. Penney, et al., Phys. Rev. Lett. **68**, 2664 (1992).
109. H. Ohno, A. Shen, F. Matsukura, et al., Appl. Phys. Lett. **69**, 363 (1996).
110. Y. D. Park, A. T. Hanbicki, S. C. Erwin, et al., Science **295**, 651 (2002).
111. F. Tsui, L. He, L. Ma, et al., Phys. Rev. Lett. **91**, 177203 (2003).
112. C. Liu, F. Yun, and H. Morkoc, J. Mater. Sci. **16**, 555 (2005).
113. S. J. Potashnik, K. C. Ku, S. H. Chun, et al., Appl. Phys. Lett. **79**, 1495 (2001).
114. K. C. Ku, S. J. Potashnik, R. F. Wang, et al., Appl. Phys. Lett. **82**, 2302 (2003).
115. F. Matsukura, H. Ohno, A. Shen, et al., Phys. Rev. B (Rapid Commun.) **57**, 2037 (1998).
116. C. Zener, Phys. Rev. **81**, 440 (1951).
117. T. Dietl, H. Ohno, F. Matsukura, et al., Science **287**, 1019 (2000).
118. T. Dietl, H. Ohno, and F. Matsukura, Phys. Rev. B **63**, 195205 (2001).
119. K. M. Yu, W. Walukiewicz, T. Wojtowicz, et al., Phys. Rev. B (Rapid Commun.) **65**, 201303 (2002).
120. S. C. Erwin and A. G. Petukhov, Phys. Rev. Lett. **89**, 227201 (2002).
121. B. Beschoten, P. A. Crowell, I. Malajovich, et al., Phys. Rev. Lett. **83**, 3073 (1999).
122. K. S. Burch, D. B. Shrekenhamer, E. J. Singley, et al., Phys. Rev. Lett. **97**, 087208 (2006).
123. R. K. Kawakami, E. Johnston-Halperin, L. F. Chen, et al., Appl. Phys. Lett. **77**, 2379 (2000).
124. A. M. Nazmul, S. Sugahara, and M. Tanaka, Phys. Rev. B (Rapid Commun.) **67**, 241308 (2003).
125. D. Chiba, F. Matsukura, and H. Ohno, Appl. Phys. Lett. **89**, 162505 (2006).
126. A. M. Nazmul, S. Kobayashi, S. Suguhara, et al., Phys. E **21**, 937 (2004).
127. S. A. Crooker, J. J. Baumberg, F. Flack, et al., Phys. Rev. Lett. **77**, 2814 (1996).
128. J. M. Kikkawa and D. D. Awschalom, Nature **397**, 139 (1999).
129. Y. Kato, R. C. Myers, A. C. Gossard, et al., Nature **427**, 50 (2004).
130. G. Salis, Y. Kato, K. Ensslin, et al., Nature **414**, 619 (2001).
131. J. A. Gupta, R. Knobel, N. Samarth, et al., Science **292**, 2458 (2001).
132. F. Meier and B. P. Zachachrenya, *Optical Orientation, Modern Problems in Condensed Matter Science* (Elsevier Science Ltd. North-Holland, Amsterdam, 1984).
133. B. Beschoten, E. Johnston-Halperin, D. K. Young, et al., Phys. Rev. B (Rapid Commun.) **63**, 121202 (2001).
134. Y. A. Bychkov and E. I. Rashba, J. Phys. C **17**, 6039 (1984).
135. G. Dresselhaus, Phys. Rev. **100**, 580 (1955).
136. S. Datta and B. Das, Appl. Phys. Lett. **56**, 665 (1990).
137. S. A. Crooker and D. L. Smith, Phys. Rev. Lett. **94**, 236601 (2005).
138. G. Schmidt, D. Ferrand, L. W. Molenkamp, et al., Phys. Rev. B (Rapid Commun.) **62**, 4790 (2000).
139. E. I. Rashba, Phys. Rev. B (Rapid Commun.) **62**, 16267 (2000).
140. A. Fert and H. Jaffres, Phys. Rev. B **64**, 184420 (2001).
141. H. J. Zhu, M. Ramsteiner, H. Kostial, et al., Phys. Rev. Lett. **87**, 016601 (2001).
142. A. T. Hanbicki, B. T. Jonker, G. Itskos, et al., Appl. Phys. Lett. **80**, 1240 (2002).
143. J. Strand, B. D. Schultz, A. F. Isakovic, et al., Phys. Rev. Lett. **91**, 036602 (2003).
144. A. T. Hanbicki, O. M. J. van't Erve, R. Magno, et al., Appl. Phys. Lett. **82**, 4092 (2003).
145. V. F. Motsnyi, J. De Boeck, J. Das, et al., Appl. Phys. Lett. **81**, 265 (2002).
146. X. Jiang, R. Wang, R. M. Shelby, et al., Phys. Rev. Lett. **94**, 056601 (2005).
147. H. Dery and L. J. Sham, Phys. Rev. Lett. **98**, 046602 (2007).
148. C. Ciuti, J. P. McGuire, and L. J. Sham, Phys. Rev. Lett. **89**, 156601 (2002).
149. J. Stephens, J. Berezovsky, J. P. McGuire, et al., Phys. Rev. Lett. **93**, 097602 (2004).
150. S. A. Crooker, M. Furis, X. Lou, et al., Science **309**, 2191 (2005).
151. S. O. Valenzuela, D. J. Monsma, C. M. Marcus, et al., Phys. Rev. Lett. **94**, 196601 (2005).
152. C. Ciuti, J. P. McGuire, and L. J. Sham, Appl. Phys. Lett. **81**, 4781 (2002).
153. H. Dery, P. Dalal, L. Cywinski, et al., Nature **447**, 573 (2007).

154. T. Schapers, J. Nitta, H. B. Heersche, et al., Phys. Rev. B **64**, 125314 (2000).
155. M. Johnson and R. H. Silsbee, Phys. Rev. B **37**, 5312 (1988).
156. S. O. Valenzuela and M. Tinkham, Appl. Phys. Lett. **85**, 5914 (2004).
157. Y. Ji, A. Hoffman, J. Pearson, et al., Appl. Phys. Lett. **88**, 052509 (2006).
158. I. Appelbaum, B. Huang, and D. J. Monsma, Nature **447**, 295 (2007).
159. Y. K. Kato, R. C. Myers, A. C. Gossard, et al., Science **306**, 1910 (2004).
160. J. Wunderlich, B. Kaestner, J. Sinova, et al., Phys. Rev. Lett. **94**, 047204 (2005).
161. S. O. Valenzuela and M. Tinkham, Nature **442**, 176 (2006).
162. T. Kimura, Y. Otani, T. Sato, et al., Phys. Rev. Lett. **98**, 156601 (2007).
163. E. Saitoh, M. Ueda, H. Miyajima, et al., Appl. Phys. Lett. **88**, 182509 (2006).
164. J. E. Hirsch, Phys. Rev. Lett. **83**, 1834 (1999).
165. S. Zhang, Phys. Rev. Lett. **85**, 393 (2000).
166. H.-A. Engel, B. I. Halperin, and E. I. Rashba, Phys. Rev. Lett. **95**, 166605 (2005).
167. W.-K. Tse and S. Das Sarma, Phys. Rev. Lett. **96**, 056601 (2006).
168. S. Murakami, N. Nagaosa, and S.-C. Zhang, Science **301**, 1348 (2003).
169. J. Sinova, D. Culcer, Q. Niu, et al., Phys. Rev. Lett. **92**, 126603 (2004).

Chapter 6
Transport in Nanostructures

Stephen M. Goodnick

6.1 Introduction

The past decade has witnessed an enormous growth of a quite diverse set of multidisciplinary science and engineering disciplines broadly falling under an umbrella called 'nanotechnology'. Nanotechnology literally implies technology at nanometer scale dimensions (10^{-9} m). From that standpoint, nanotechnology is not a recent phenomenon; nanostructured materials have been used for centuries to enhance the properties of tools, ceramics, building materials, etc. (tempered steel used for sword making is a good example). However, the historical applications of nanotechnology were purely empirical, with no underlying knowledge of the nanoscale material structure. In contrast, the current nanotechnology revolution is driven by and large by our ability to probe, analyze, and manipulate matter at this size scale. The transition from the 'macro' to 'micro' to 'nano' is not abrupt, but occurs smoothly over multiple length scales. As a result, there is quite a bit of ambiguity, in what is truly 'nanotechnology' as opposed to microelectronics, micromachining, cellular biology, etc. Somewhat arbitrarily, we define *nanometer scale* to characteristic feature sizes on the order of 100 nm, or less in terms of the separation of the micro- and nano-worlds.

Nanoelectronics generally refers to nanometer scale devices, circuits, and architectures impacting continued scaling of information processing systems, including communication and sensor systems, as well as providing an interface between the electronic and biological worlds. The present attention on nanotechnology and nanoelectronics has been driven from the top down by the continued scaling of semiconductor device dimensions into the nanometer scale regime, as discussed in more detail below. It is predicted that the scaling down of dimensions in present semiconductor technologies will continue for the next 8–10 years, until a hard limit of Moore's Law is finally reached due to manufacturability, or

S.M. Goodnick
Department of Electrical Engineering, Arizona State University, P.O. Box 875706, Tempe, AZ 85287-5706, USA
e-mail: Stephen.Goodnick@asu.edu

A. Korkin, F. Rosei (eds.), *Nanoelectronics and Photonics*,
DOI: 10.1007/978-0-387-76499-3_6, © Springer Science+Business Media, LLC 2008

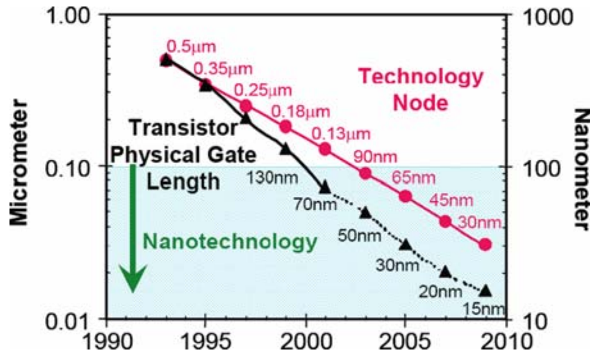


Fig. 6.1 Scaling of semiconductor device dimensions as a function of time. The upper curve is the so-called technology node, while the lower curve (*triangles*) represents the present and projected physical gate length of the corresponding transistor technology. Reprinted with permission from R. Chau, presented at the 2005 IEEE VLSI-TSA International Symposium on VLSI Technology, Hsinchu, Taiwan, April, 2005

finally due to reaching atomic dimensions themselves. Figure 6.1 illustrates the decrease in feature size for complementary metal oxide semiconductor (CMOS) transistors (the gate length of a transistor being the critical dimension), with time corresponding to the scaling of semiconductor technology.

At the end of the roadmap for CMOS technology (10–15 nm gate lengths based on present projections), it will be necessary for radical new technologies to be introduced if continued progress in reducing device dimensions and increasing chip density is to be maintained. This 'end of the roadmap' implies that industry faces an enormous challenge of developing commercially viable nanoscale chip technologies within the next 10 years. Fundamental advances are needed in new switching mechanisms, new computing paradigms realized from locally connected architectures such as cellular non-linear networks (CNN), new ways to design for fault tolerance, new methods to achieve low power circuit design, and new methods for testing very dense and highly integrated nanoscale systems-on-a-chip.

From the molecular scale side or 'bottom up', the nanotechnology 'revolution' has been enabled by remarkable advances in atomic scale probes and nanofabrication tools. Structures and images at the atomic scale have been made possible by the invention of the scanning tunneling microscope (STM) and the associated atomic force microscope (AFM) [1]. Such scanning probe microscopy (SPM) techniques allow atomic scale resolution imaging of atomic positions, spectroscopic features, and positioning of atoms on a surface. Concurrently, there have been significant advances in the synthesis and control of self-assembled systems, semiconductor nanowires, molecular wires, and novel states of carbon such as fullerenes and carbon nanotubes. These advances have led to an explosion of scientific breakthroughs in studying the properties of individual molecular structures with potential application as components of

molecular electronic (moletronic) devices and circuits. As discussed later, such bottom up technology for novel materials growth and potential device fabrication is more closely akin to the self-assembly and complex templated structure formation found in biological systems, i.e., biomimetic structures.

6.1.1 Issues in Semiconductor Device Scaling

As the density of integrated circuits continues to increase, there is a resulting need to shrink the dimensions of the individual devices of which they are comprised. Smaller circuit dimensions reduce the overall die area, thus allowing for more transistors on a single die without negatively impacting the cost of manufacturing. As semiconductor feature sizes shrink into the nanometer scale regime, device behavior becomes increasingly complicated as new physical phenomena at short dimensions occur, and limitations in material properties are reached. In addition to the problems related to the actual operation of ultrasmall devices, the reduced feature sizes require more complicated and time-consuming manufacturing processes.

For silicon MOSFETs, in conventional device scaling, the device size is scaled in all dimensions, resulting in smaller oxide thickness, junction depth, channel length, channel width, and isolation spacing. Figure 6.2 shows the scaling of planar MOS transistor technology from a series of electron micrographs of successive sub-100 nm gate length devices. Advances in lithography

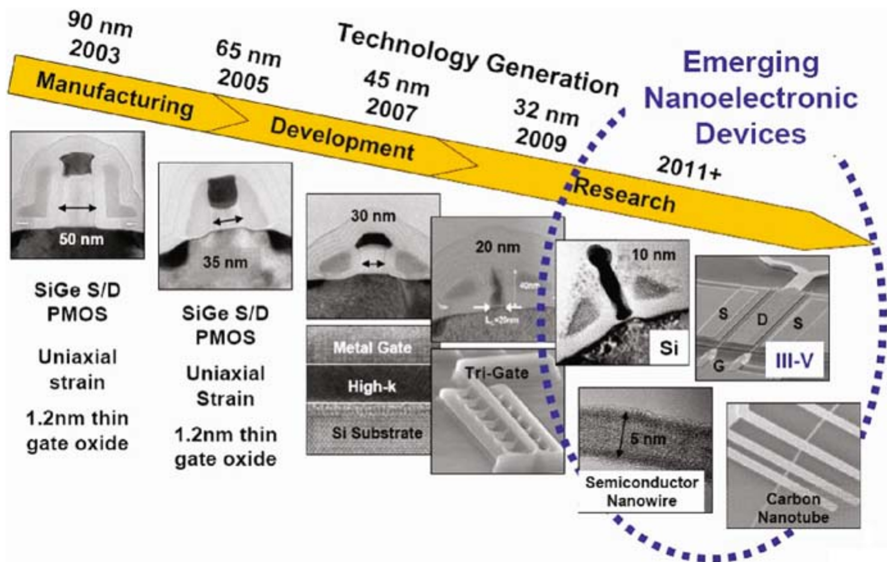


Fig. 6.2 Scaling of successive generations of MOSFETs into the nanoscale regime and emerging nanoelectronic devices. Reprinted with permission from R. Chau, presented at INFOS 2005

have driven device dimensions to the deep-submicrometer range, where gate lengths are drawn at 0.1 μm and below. The Semiconductor Industry Association (SIA) projects that by the end of 2009, leading edge production devices will employ 25 nm gate lengths and have oxide thickness of 1.5 nm, or less [2]. In fact, laboratory MOSFET devices with gate lengths down to 15 nm have been reported, which exhibit excellent *I–V* characteristics [3]. Beyond that, there has been extensive work over the past decade related to nanoelectronic or quantum scale devices which operate on very different principles from conventional MOSFET devices, but may allow the continued scaling beyond the end of the current scaling roadmap [4]. This trend has been motivated by the fact that the performance of the scaled device in the 25 nm regime is itself problematic, as discussed below.

For example, to enhance device performance, the gate oxide thickness has to be aggressively scaled. However, as the gate oxide thickness approaches 1 nm through scaling, tunneling through the gate oxide results in unacceptably large off-state currents, dramatically increasing quiescent power consumption [5], and rendering the device impractical for analog applications due to unacceptable noise levels. Another consequence of scaling is that the stack of layered materials that comprise electronic devices is becoming more like a continuum of interfaces rather than a stack of bulk thin films. Therefore, topology effects arising from surface(interface)-to-surface(interface) interactions now dominate the formation of potential barriers at interfaces. The interface inhomogeneity effects include morphological and compositional inhomogeneities. Morphological inhomogeneities, typically manifested as atomic scale roughness, are often responsible for increased leakage currents in MOSFET gates. Fluctuations in the elemental distribution are expressions of compositional inhomogeneities. For finite dimensions and number of atoms, interface domains cannot be represented as superpositions of a few homogeneous thin film regions. Instead, the challenge of characterizing this complex system requires accurate atomic level information about the three-dimensional (3D) structure, geometry, and composition of atomic scale interfaces.

Yet another issue that will pose serious problems on the operation of future ultrasmall devices is related to the substrate doping used to gain control of the electrophysical properties of the semiconductor and the operational parameters of electronic devices by control of the type, concentration, and distribution of impurities. The distribution of dopants is traditionally treated as continuum in semiconductor physics, which implies the following: (a) the number of impurity atoms is small as compared to the total number of atoms in the semiconductor matrix and (b) the impurity atoms distribution is statistically uniform, while the position of an individual atom in the lattice is not defined, e.g., is random. The assumption of statistical uniformity requires large number of atoms, which is not the case in, for example, a 25 nm MOSFET device in which one has less than 100 dopant atoms in the junction region. In these future ultrasmall devices, the number and location of each dopant atom will play an important role in determining the overall device behavior. The challenge of precisely placing

small number of dopants may represent an insurmountable barrier, which could end conventional MOSFET scaling.

Quantum mechanical effects due to spatial quantization in the device channel region may play an important role in the operation of nanoscale devices. Quantization affects both the charge distribution in the channel (and hence the capacitance) and transport. Quantum confinement results in a setback of charge from the oxide–semiconductor interface, which adds to the effective thickness of the gate oxide, which for a 25 nm MOSFET device is already on the order of 1 nm. This leads to a decrease of effective gate capacitance and a shift in the threshold voltage. Another issue affecting device performance is carrier transport along the channel. Because of the two-dimensional (2D) confinement of carriers in the channel, the carrier mobility is different from the 3D case, as discussed in Section 6.3. Theoretically speaking, the 2D mobility should be larger than its 3D counterpart due to reduced density of states function, i.e., reduced number of final states the carriers can scatter into, although surface scattering in turn may reduce the mobility. In the limit of very short gate length devices, carriers should be almost ballistic, which makes the issue of scattering less relevant, but still an important parameter in device performance.

6.1.2 Non-classical and Quantum Effect Devices

To fabricate devices beyond current scaling limits, CMOS technology is rapidly moving toward quasi-3D structures such as dual-gate, tri-gate, and Fin-FET structures [6], in which the active channel is increasingly a nanowire or nanotube rather than bulk region. Figure 6.3 illustrates a schematic of a Fin-FET device, and the corresponding electron micrograph of a multi-gate Fin-FET

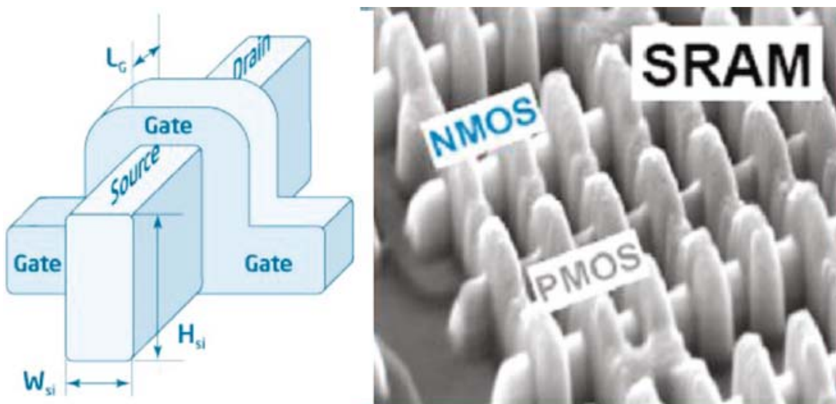


Fig. 6.3 Non-classical device structures. *Left:* a schematic of a FinFET; *right*: an SEM photo of multileg NMOS and PMOS FinFET structures forming part of an SRAM. Reprinted with permission from R. Chau

architecture. Here the heavily doped Si substrate is replaced by a Si on insulator (SOI) substrate, with a buried oxide layer (BOX) supporting the thin Si channel. The fin gate wraps around the side of the channel. Such 3D gate structures are needed to maintain charge control in the channel, as channel lengths scale toward nanometer dimensions.

Beyond field effect transistors, there have been numerous studies over the past two decades of alternatives to classical CMOS at the nanoscale. As discussed in more detail in Section 6.2, when the critical dimensions become shorter than the phase coherence length of electrons, the quantum mechanical wave nature of electrons becomes increasingly apparent, leading to phenomena such as interference, tunneling, and quantization of energy and momentum as discussed earlier. Indeed, for a one-dimensional (1D) wire, the system may be considered a waveguide with 'modes', each with a conductance less than or equal to a fundamental constant $2e^2/h$, discussed in detail in Section 6.4.2. Such quantization of conductance was first measured in split-gate field effect transistors at low temperatures [7, 8], but manifestations of quantized conductance appear in many transport phenomena such as universal conductance fluctuations [9] and the quantum Hall effect [10]. While various early schemes were proposed for quantum interference devices based on analogies to passive microwave structures (see, for example, [11, 12, 13]), most suffer from difficulty in control of the desired waveguide behavior in the presence of unintentional disorder. This disorder can arise from the discrete impurity effects discussed earlier, as well as the necessity for process control at true nanometer scale dimensions. More recently, promising results have been obtained on ballistic Y-branch structures [14], where non-linear switching behavior has been demonstrated even at room temperature [15].

In the previous section, we discussed the role of discrete impurities as an undesirable element in the performance of nanoscale FETs. However, the discrete nature of charge in individual electrons, and control of charge motion of single electrons, has in fact been the basis of a great deal of research in single electron devices and circuits (see, for example, [16]), as discussed in more detail in Section 6.5. The understanding of single electron behavior is most easily provided in terms of the capacitance, C, of a small tunnel junction, and the corresponding change in electrostatic energy, $E = e^2/2C$, when an electron tunnels from one side to the other. When physical dimensions are sufficiently small, the corresponding capacitance (which is a geometrical quantity in general) is correspondingly small, so that the change in energy is greater than the thermal energy, resulting in the possibility of a 'Coulomb blockade', or suppression of tunnel conductance due to the necessity to overcome this electrostatic energy. This Coulomb blockade effect allows the experimental control of electrons to tunnel one by one across a junction in response to a control gate bias (see, for example, [4, 17]). Single electron transistors [18], turnstiles [19, 20], and pumps [21] have been demonstrated, even at room temperature [22]. Computer-aided modeling tools have even been developed based on Monte Carlo simulation of charge tunneling across arrays of junctions, to facilitate the

design of single electron circuits [23]. As in the case of quantum interference devices, the present-day difficulties arise from fluctuations due to random charges and other inhomogenieties, as well as the difficulty in realizing lithographically defined structures with sufficiently small dimensions to have charging energies approaching kT and above.

There has been rapid progress in realizing functional nanoscale electronic devices based on self-assembled structures such as semiconductor nanowires (NWs) [24] and carbon nanotubes (CNTs) [25]. Semiconductor nanowires have been studied over the past decade in terms of their transport properties [4], and for nanodevice applications such as resonant tunneling diodes [26], single electron transistors [27, 28], and field effect structures [24]. Recently, there has been a dramatic increase in interest in NWs due to the demonstration of directed self-assembly of NWs via in situ epitaxial growth [29, 30]. Such semiconductor NWs can be elemental (Si,Ge) or III–V semiconductors, where it has been demonstrated that such wires may be controllably doped during growth [31], and abrupt compositional changes forming high-quality 1D heterojunctions can be achieved [32, 33]. A variety of different device technologies have been achieved with self-assembled nanowire growth, as discussed later.

Likewise, CNTs have received considerable attention due to the ability to synthesize NTs with metallic, semiconducting, and insulating behavior, depending primarily on the chirality (i.e., how the graphite sheets forming the structure of the CNT wrap around and join themselves) [34]. Semiconducting CNTs may be doped to realize n-type and p-type semiconducting wires, which are the basis of a number of demonstrations of transistors, logic circuits, and sensors.

Summarizing the above discussion, there are a variety of new phenomena that become important as device dimensions scale to the nanoscale and beyond. These include the following:

- Quantum confinement – small dimensions lead to quantum confinement and associated quantization of motion leading to discrete energy levels.
- Quantum interference – at dimensions smaller than the phase coherence length, the wave-like behavior of particles manifests itself, leading to reflection, refraction, tunneling, and other non-classical wave-like behavior.
- Phase coherent transport – at dimensions smaller than the mean free path for scattering, transport is ballistic rather than diffusive.
- Single electron effects – for small structures, the discrete nature of charge itself is important, and the associate energy for transfer of charge is non-negligible compared to the total energy of the system.

The rest of this review addresses these topics individually. Section 6.2 addresses the role of length scale in the transition from the behavior of 'classical' devices to quantum effect devices. Section 6.3 looks at the role of quantum confinement on transport in reduced dimensionality systems, particularly in quantum wells and quantum wires. Section 6.4 then addresses the transition from diffusive to ballistic transport, and phenomena associated with the

transmission and reflection of electrons as wave-like objects rather than particles. Finally, Section 6.5 looks at the formation of artificial molecular structures and the effect on transport of single electron tunneling.

6.2 Overview of Electronic Transport in Nanoscale Systems

6.2.1 Electronic Transport in Semiconductors

The subject of electronic transport in semiconductors and in solids in general is a very old problem, which has been well studied over the past 75 years. A general overview is given elsewhere (see, for example, [35]). Transport is an inherently non-equilibrium phenomena, where the role of dissipation and the coupling to the environment play a crucial role. External forces which drive the system out of equilibrium may be electromagnetic in origin, such as the electric fields associated with an applied DC bias, or the excitations of electrons from their ground to excited states due to high-frequency optical excitation. Alternately, electrochemical potentials, thermal gradients, etc., may also provide the drive for electronic transport and its external manifestation in terms of macroscopic currents and voltages.

Electronic transport at its most fundamental level requires a full many-body quantum mechanical description going beyond the usual ground state descriptions of solids used in ab initio calculations of the electronic states. Clearly, a full many particle description of transport including the real number of particles in both the device, its contact to the external environment, and the external environment itself, is beyond the ability of any computational platform in the foreseeable future. Hence, successive levels of approximation that sacrifice information about the system and the exact nature of transport are necessary in any sort of realistic description of transport. Figure 6.4 illustrates the hierarchy of transport approaches used in describing electronic transport in semiconductors, metals, and molecular systems. At the bottom is the exact solution of the N-body quantum mechanical problem which is computationally intractable except for small numbers of particles (less than 100). To treat the many-body problem, some sort of mean-field approximation is necessary which transforms the problem into an effective one-electron problem. Non-equilibrium Green function methods are currently popular at the next level of approximation as they contain retain important correlations in space and time, which are believed to be important at the nanoscale. Above this are quantum kinetic approaches in terms of the Liouville–von Neumann equation of motion for the density matrix, or Wigner distribution approaches that contain quantum correlations but retain the form of semi-classical approaches in terms of the distribution function. In going from the quantum to the classical description of charge transport, information concerning the phase of the electron and its non-local behavior is lost, and electronic transport is treated in

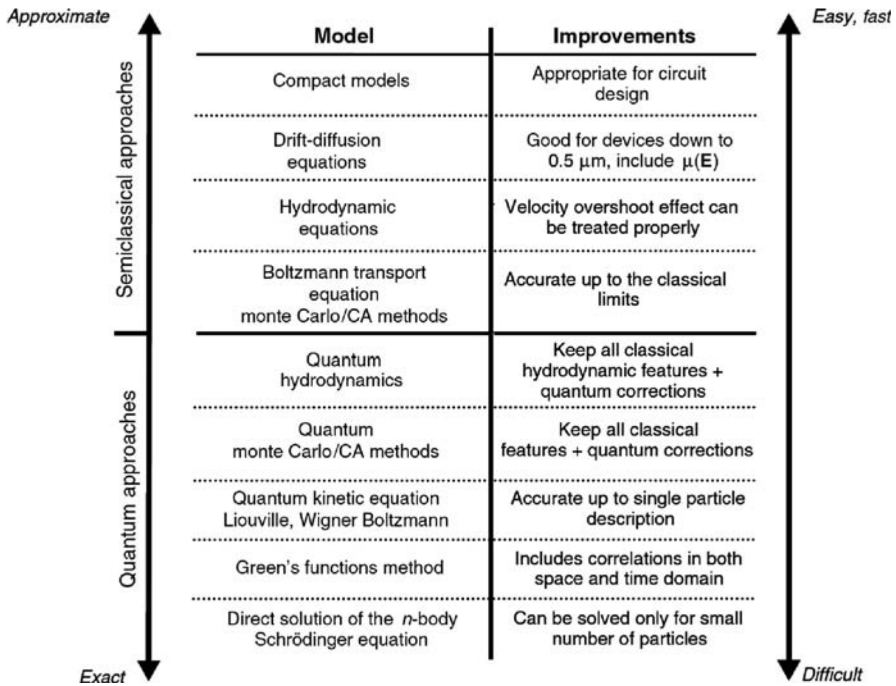


Fig. 6.4 Hierarchy of transport approaches used in the description of electronic transport in semiconductors [36]

terms of a purely particle framework. This is the level of the Boltzmann transport equation (BTE), which represents a kinetic equation describing the time evolution of the distribution function describing both the position and the momentum of the particle, and has been the primary framework for describing transport in semiconductors and semiconductor devices with microscale and above dimensions. There are then approximations to the BTE, given by moment expansions of the BTE which lead to the hydrodynamic, the drift-diffusion, and relaxation time approximation approaches to transport (the latter given the Drude form of conductivity). Finally, at an empirical level are non-linear circuit models for device behavior suitable for circuit simulation in the so-called compact models.

One interesting aspect of transport in nanostructure systems is that the characteristic length scales span the transition from classical to quantum transport. Hence a single description in the hierarchy of Fig. 6.1 may not be sufficient, or may be overly cumbersome for providing the correct physics of device operation. Depending on certain critical length scales, discussed in the next section, transport may be semi-classical or purely quantum, or even more difficult, a mixture of the two in which the effects of decoherence and dissipation play important roles, while at the same time, quantum effects still dominant.

6.2.2 Transport in Nanoscale Systems

As mentioned above, transport in nanoscale systems is often a function of the characteristic length scale associated with the motion of carriers. To understand this notion better, consider a prototypical nanodevice illustrated in Fig. 6.5. The 'device' is coupled to two contacts, left and right, which serve as a source and sink (drain) for electrons. Here the contacts are drawn as metallic-like reservoirs, characterized by chemical potentials μ_s and μ_D, and are separated by an external bias, $qV_A = \mu_s - \mu_D$. The current flowing through the device is then a property of the chemical potential difference and the transmission properties of the active region itself. A separate gate electrode serves to change the transmission properties of the active region, and hence modulates the current. This separation of a nanodevice into ideal injecting and extracting contacts, and an active region which limits the transport of charge, is a common way of visualizing the transport properties of nanoscale systems. However, it clearly has limitations, the contacts themselves are really part of the active system, and are driven out of equilibrium due to current flow, as well as coupling strongly to the active region through the long-range Coulomb interaction of charge carriers.

The nature of transport in a nanodevice such as that illustrated in Fig. 6.5 depends on the characteristic length scales of the active region of the device, L. Figure 6.6 illustrates the active region of this nanodevice in terms of a conductor of length L, and width W. The mean free path between collisions is designated l, while the length scale over which quantum coherence is preserved (the phase breaking length) is designated l_φ. The latter is often associated with the inelastic mean free path, or the distance between dissipative scattering events where the inelastic coupling to the environment is associated with quantum mechanical phase breaking. Figure 6.6a corresponds to the case in which both L and W are much larger than both the elastic and inelastic mean free paths. Here transport is purely diffusive, and the system behaves essentially as a semi-classical metal or semiconductor governed by the BTE in the hierarchy of Fig. 6.4. In Fig. 6.6b,

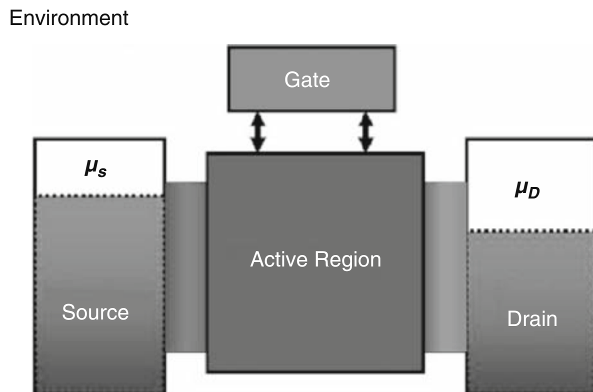


Fig. 6.5 Schematic of a generic nanoelectronic device consisting of sources and sinks for charge carriers (the source and drain contacts), and a 'gate' which controls the transfer characteristics of the active region

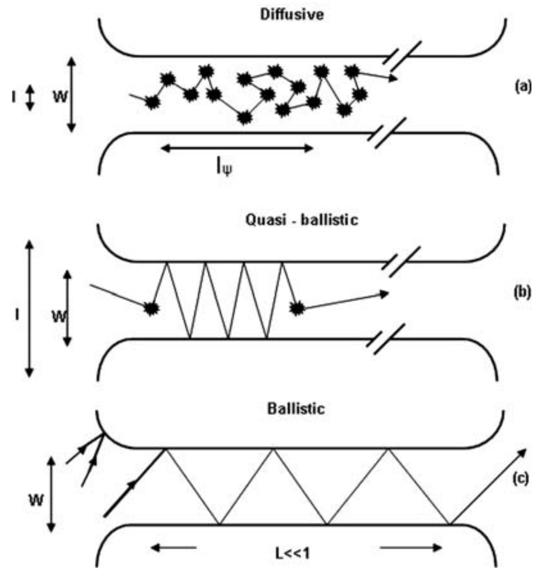


Fig. 6.6 Illustration of the effect of length scale on transport in nanoscale systems. L and W represent the length and width of a nanoscale conductor. The elastic mean free path is designated l, and the phase coherence length is l_{φ}

the width, W, is smaller than the characteristic mean free path, while the length, L, is still much longer. This regime corresponds to the case of a quantum confined system, in which the motion of carriers is quantized in one dimension, but essentially behaves as a diffusive conductor in the other directions. Quasi-2D and quasi-1D systems such as those discussed in Section 6.3 correspond to this case. Finally, when both L and W are shorter than the elastic and inelastic mean free paths, the system is purely ballistic, and the motion of charge is governed by the wave-like behavior of the particle and its reflection and transmission properties through the structure.

6.3 Diffusive Transport in Quantum Confined Systems

6.3.1 Semi-classical Boltzmann Transport Equation

As mentioned above, the classical description of charge transport is given by the BTE in the hierarchy of Fig. 6.4. The BTE is an integral–differential kinetic equation of motion for the probability distribution function for particles in the 6D phase space of position and (crystal) momentum:

$$\frac{\partial f(\mathbf{r},\mathbf{k},t)}{\partial t} + \frac{1}{\hbar}\nabla_{\mathbf{k}}E(\mathbf{k})\cdot\nabla_{\mathbf{r}}f(\mathbf{r},\mathbf{k},t) + \frac{\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}}f(\mathbf{r},\mathbf{k},t) = \left.\frac{\partial f(\mathbf{r},\mathbf{k},t)}{\partial t}\right|_{\mathrm{Coll}}, \quad (6.1)$$

where $f(\mathbf{r}, \mathbf{k}, t)$ is the one-particle distribution function. The right hand side is the rate of change of the distribution function due to randomizing collisions and is an integral over the in-scattering and the out-scattering terms in momentum (wavevector) space. Once $f(\mathbf{r}, \mathbf{k}, t)$ is known, physical observables, such as average velocity or current, are found from averages of f. Equation (6.1) is semi-classical in the sense that particles are treated as having distinct position and momentum in violation of the quantum uncertainty relations, yet their dynamics and scattering processes are treated quantum mechanically through the electronic band structure (and the use of time-dependent perturbation theory.

The BTE itself is an approximation to the underlying many-body classical Liouville equation, and quantum mechanically by the Liouville–von Neumann equation of motion for the density matrix. The main approximations inherent in the BTE are the assumption of instantaneous scattering processes in space and time, the Markov nature of scattering processes (i.e., that they are uncorrelated with the prior scattering events), and the neglect of multi-particle correlations (i.e., that the system may be characterized by a single particle distribution function). The inclusion of quantum effects such as particle interference, tunneling which take one further down the hierarchy of Fig. 6.4 is more problematic in the semi-classical *Ansatz*, and is an active area of research today as device dimensions approach the quantum regime.

Free carriers (electrons and holes) interact with the crystal and with each other through a variety of scattering processes which relax the energy and momentum of the particle. Based on first-order, time-dependent perturbation theory, the transition rate from an initial state $\mathbf{k}$ in band n to a final state $\mathbf{k}'$ in band m for the jth scattering mechanism is given by Fermi's Golden rule [37]:

$$\Gamma_j[n, \mathbf{k}; m, \mathbf{k}'] = \frac{2\pi}{\hbar} | < m, \mathbf{k}' | V_j(\mathbf{r}) | n, \mathbf{k} > |^2 \delta(E_{\mathbf{k}'} - E_{\mathbf{k}} \mp \hbar\omega), \tag{6.2}$$

where $V_j(\mathbf{r})$ is the scattering potential of this process and $E_{\mathbf{k}}$ and $E_{\mathbf{k}}'$ are the initial and final state energies of the particle. The delta function results in conservation of energy for long times after the collision is over, with $\hbar\omega$ the energy absorbed (upper sign) or emitted (lower sign) during the process. Scattering rates calculated by Fermi's Golden rule above are typically used in Monte Carlo device simulation as well as in simulation of ultrafast processes. The total rate used to generate the free flight in Eq. (6.2), discussed in the previous section, is then given by

$$\Gamma_j[n, \mathbf{k}] = \frac{2\pi}{\hbar} \sum_{m,\mathbf{k}'} | < m, \mathbf{k}' | V_j(\mathbf{r}) | n, \mathbf{k} > |^2 \delta(E_{\mathbf{k}'} - E_{\mathbf{k}} \mp \hbar\omega). \tag{6.3}$$

There are major limitations to the use of the Golden rule due to effects such as *collision broadening* and *finite collision duration time*. The energy-conserving delta function is only valid asymptotically for times long after the collision is

complete. The broadening in the final state energy is given roughly by $\Delta E \approx \hbar/\tau$, where τ is the time after the collision, which implies that the normal $E(\mathbf{k})$ relation is only recovered at long times. Beyond this, there is still the problem of dealing with the quantum mechanical phase coherence of carriers, which are important in nanodevices, as discussed in Section 6.4.

Figure 6.7 shows a taxonomy of various scattering mechanisms occurring in a typical semiconductor system. These are roughly device in terms of elastic processes (defect scattering), dissipative processes (lattice scattering), and ineleastic intercarrier scattering processes. Defect scattering occurs due to static defects in the otherwise perfect crystal lattice, the strongest of which is usually ionized impurity scattering due to the long-range Coulomb interaction. Other defects such as vacancies and dislocations can be effects scattering processes depending whether they are charged or not. Alloy scattering occurs in semiconductor alloys such as SiGe or ternary alloys such as InGaAs, due to the random occurrence of the component species on a particular lattice site. Lattice scattering is associated with the electron–phonon interaction, which is usually described with a deformation potential approach, with either acoustic or optical modes of the crystal. In polar semiconductors (all III–V and II–VI compounds for example), the polar optical interaction due to the fluctuating dipole moment of the charged cation–anion pair (the Fröhlich interaction), is quite effective and limits the mobility of intrinsic materials at room temperature. Similarly, piezoelectric coupling with acoustic modes in polar materials can be an effective scattering process as well. The various types of deformation scattering are often categorized as intravalley versus intervalley, to distinguish scattering process that take carriers from the minimum of one conduction band in k-space to the minimum of a different conduction band.

There are various methods for solving the BTE under certain simplifying assumptions such as the relaxation time approximation (leading to the Drude

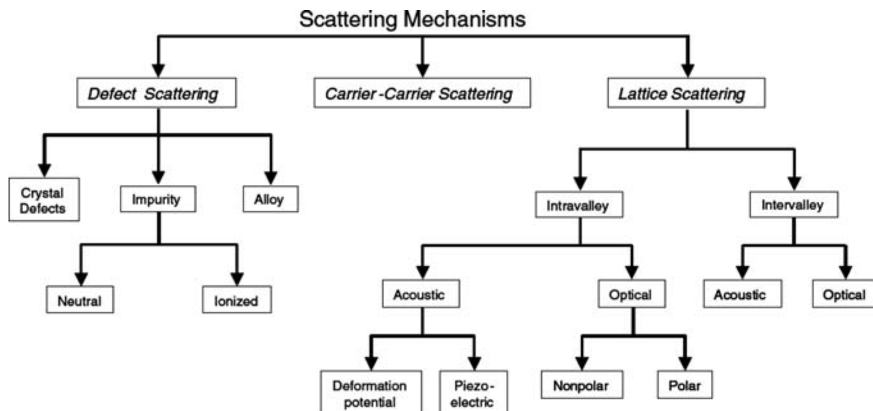


Fig. 6.7 Scattering processes in a typical semiconductor

conductivity at low temperatures), or moment expansions of the BTE. A popular method of solution is the Ensemble Monte Carlo technique [38], in which the motion in time of an ensemble of pseudo-particles is simulated according to their deterministic free flights, and random scattering events generated using a random number generator and the appropriate transition rates calculated from Eqs. (6.2) and (6.3). Such simulations can include the full bandstructure and phonon dynamics of the semiconductor, as shown in Fig. 6.8, which shows the calculated bandstructure for wurzite GaN, the corresponding scattering rates for various processes as a function of energy, and the

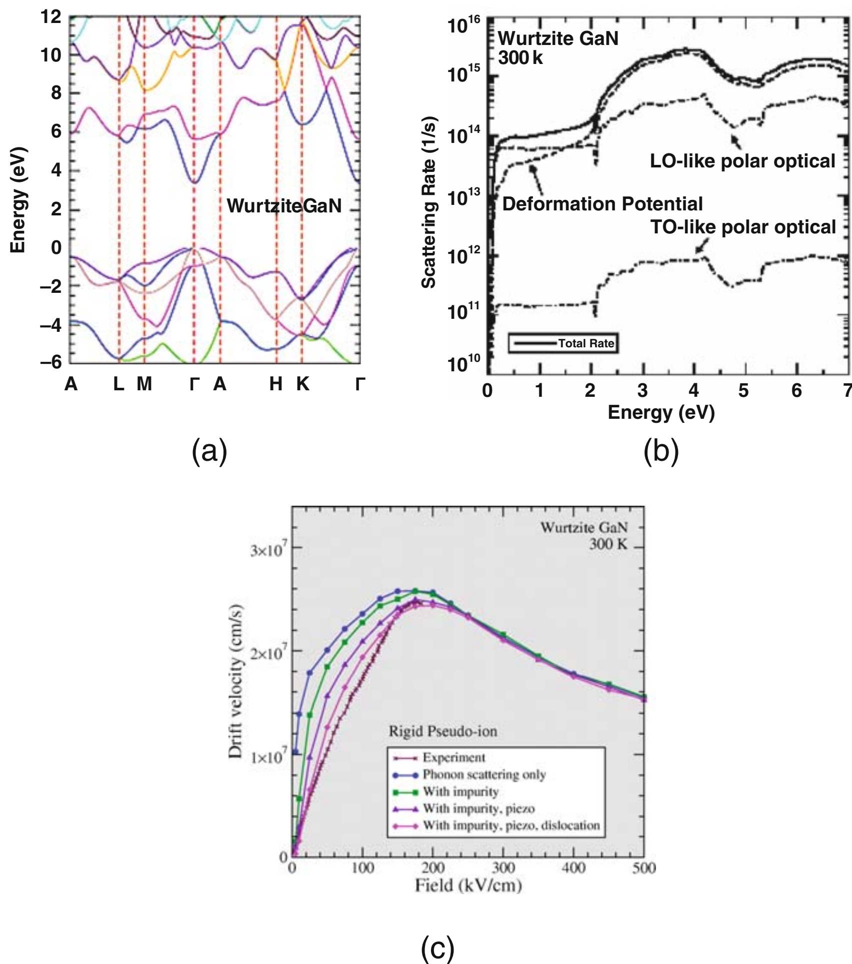


Fig. 6.8 Calculated bandstructure (**a**), rigid-ion scattering rates (**b**), and calculated velocity-field characteristics (**c**) from the CMC simulator for Wurzite GaN at 300 K. Reprinted with permission from Ref. [39], Copyright 2004, Institute of Physics Publishing (*See* Color Insert)

calculated velocity-field characteristics assuming different scattering processes in comparison to experiment [39].

6.3.2 Effects of Quantum Confinement on Transport

As discussed earlier, as the characteristic length scales decrease, quantum mechanical effects become more important, as illustrated in Fig. 6.6. One important effect is quantum confinement due to heterostructure barriers, free surfaces, electrostatic potential confinement, etc., resulting in *reduced dimensionality* of the system. Confinement in one direction only creates a *quantum well*, typically in the growth direction of a heterostructure or oxide– semiconductor system. Each of the bound state solutions of the Schrödinger equation in the confinement direction corresponds to the formation of a subband, with a localized envelope function in the confined direction and free electrons in the unconfined directions. The unconfined electrons in the lateral direction form a *quasi-2D gas*, where particles behave as free particles within each of the subbands.

If further confinement is imposed, for example, by laterally etching a heterostructure quantum well structure to from a *nanowire*, motion is confined in two directions, and free in the longitudinal axis of the nanowire. Again, one has a series of subbands formed by the bound state solutions in the confined directions, and the free electrons comprise a *quasi-1D electron gas*. Finally, if the system is confined in all three dimensions, through either artificial patterning or self-assembly, then the energy spectrum is completely discrete, and we form a *quantum dot*, or *nanocrystal* structure.

In terms of transport, we can distinguish between transport parallel or perpendicular to the confining potentials in the system. In the latter case, transport is dominated by quantum mechanical reflection and transmission and associated non-classical phenomena such as tunneling. We discuss this case in more detail in Section 6.4. In terms of transport along one of the unconfined directions of a reduced dimensional system, over long distances (longer than the elastic mean free path), transport is diffusive, and similar to the bulk case discussed in Section 6.3.1, but modified by the reduced dimensionality and scattering between subbands.

One of the main differences between transport in varying dimensionality systems is the density states, which plays a critical role in scattering theory in terms of the availability of final states to scattering into. Generally speaking, the energy dependence of the density of states for spherical, parabolic, constant energy surfaces can be written as

$$g(E) = A(E - E_i)^n, \tag{6.4}$$

where g(E) is the density of states and $n = 1/2$ for 3D, $n = 1$ for 2D, and $n = -1/2$ for a 1D system, where A is a constant depending on the effective mass. More generally, including the bound state subband energies,

$$g(E) = A \sum_{i=1} (E - E_i)^n \Theta(E - E_i), \tag{6.5}$$

where $\Theta(E - E_i)$ is the unit step function, i denotes the subband index, and n has the same meaning as above.

A schematic of the density of states in reduced dimensionality systems is shown if Fig. 6.9. The upper plot corresponds to the density of states for a quantum well system, which is constant within each subband, corresponding to the density of states of a 2D system. The density of states for a quantum wire structure is shown in the middle figure, which has a singularity at the subband edge due to the 1D density of states associated with each subband. Finally, for a quantum dot (or quasi-2D system with a magnetic field intensity corresponding to $\hbar\omega_c$), the density of states is discrete.

Scattering between subbands, or intersubband scattering, is another non-bulk-like phenomena occurring in quantum confined systems. Figure 6.10 illustrates this process for a two-subband system. 1 and 2 label the first and second subbands of the system, whose dispersion is characteristic of the free motion in the unconfined direction. Transitions between 2 and 1 are intersubband transitions, while transitions 2 to 2 or 1 to 1 are intrasubband transitions. Intersubband transitions are critical for the relaxation of exited carriers (for example, through photoexcitation, or pumped in the cavity of a laser structure) from upper states to the ground subband of the system. Since energy and momentum generally should be conserved during intersubband transitions, some types of intersubband scattering may be suppressed, for example, when

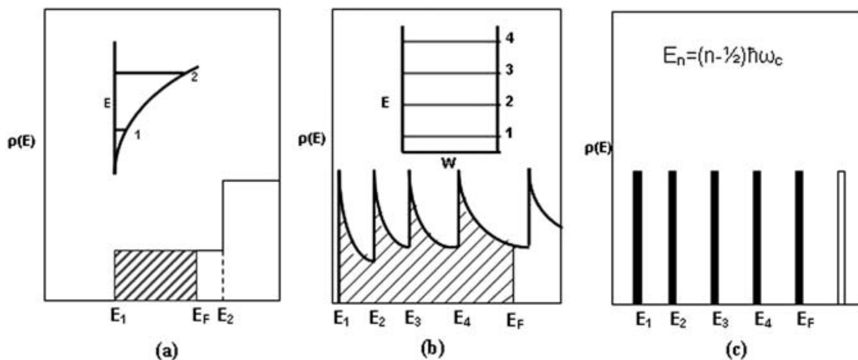


Fig. 6.9 Density of states for a quasi-2D system (**a**), quasi-1D system (**b**), and a zero-dimensional system (**c**)

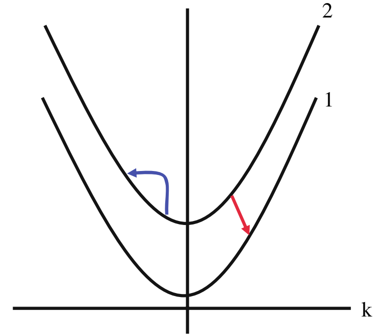


Fig. 6.10 Illustration of *intersubband scattering* versus *intrasubband scattering*

the phonon energy is larger than the intersubband energy spacing, leading to potential bottlenecks in carriers reaching the lowest subband.

In addition to the modification of the density of states, and intersubband scattering processes, there are several non-bulk scattering processes occurring in quantum confined systems which can severely limit transport, and which are primarily associated with the surfaces and interfaces associated with confinement itself: These include the following:

- Remote impurity scattering – through the process of *modulation doping*, the ionized impurities responsible for free carriers in a quantum well or nanowire are spatially separated from the conducting channel itself where free carriers reside. Hence scattering is due to ionized dopants and is greatly suppressed due to the spatial separation, allowing for very high mobilities.
- Surface roughness scattering – scattering due to fluctuations of the surface or interface associated with carrier confinement. Such random fluctuations may be a strong source of scattering, particularly when carriers are localized close to the interface, and may limit the mobility.
- Surface states and impurities – dangling bonds and impurity atoms may be present at the surface or interface which strongly couple to electrons there.
- Confined phonons – the phonon spectra itself may be modified by the presence of dielectric discontinuities in quantum confined structure, giving rise to waveguide modes and localized surface modes, whose interaction with electrons may be weaker or stronger depending on the degree of confinement.

6.3.2.1 Quasi-2D Systems

Transport in quasi-2D systems was extensively studied during the late 1960s and 1970s in connection with the observation of quantization at the Si/SiO_2 interface (see [40] for a detailed review) in metal oxide semiconductor field

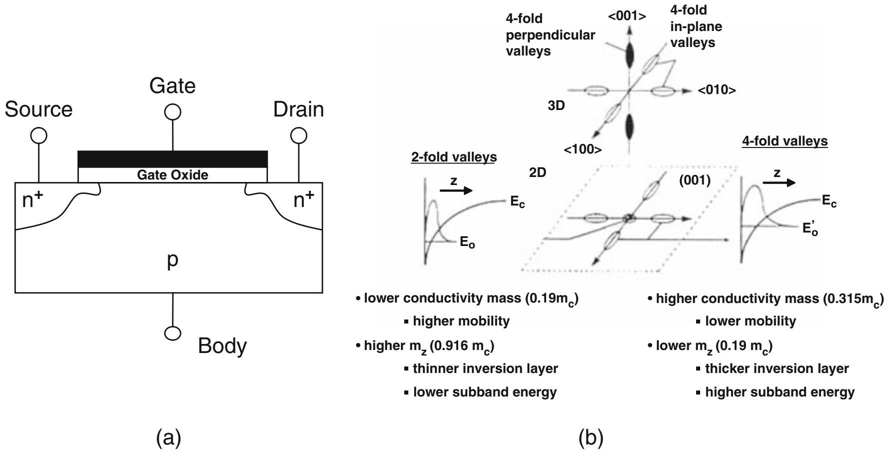


Fig. 6.11 Quasi-2D system formed at the Si/SiO_2 interface in a MOSET (metal oxide semiconductor field effect transistor) device. (**a**) Cross-section of the device. (**b**) Effect of the multi-valley conduction band structure on quantization for [100] Si (with permission from D. Vasileska, private communication)

effect transistor (MOSFET) structures. Figure 6.11a shows a cross-section of a typical MOSFET structure. With a positive potential applied to the gate, electrons are drawn to the surface forming a conducting channel between the source and drain. The electric field at the interface may be on the order of MV/cm, which forms a strong potential well, and quantum confinement, as illustrated in Fig. 6.11b. Silicon is an indirect bandgap semiconductor, with six degenerate conduction band minima close to the X point in the first Brillouin zone. Due to the symmetry breaking of the surface, this degeneracy is broken, and for the (100) Si surface shown below, two of the valleys project with one mass perpendicular to the surface, where the other four project with a small mass. Hence the two valleys with the heavier mass form the lowest state of the quantum confined system, and the lowest state of fourfold degenerate valleys begins at a higher energy.

The simplest description of the electronic states at the surface is within the single-band effective mass picture, where one may solve the separable envelope function equation:

$$\left(\frac{\hbar^2}{2m_z}\frac{\partial^2}{\partial z^2}+\frac{\hbar^2}{2m_{//}}\nabla_{\mathbf{r}}^2+V_{\mathrm{eff}}(z)\right)\psi(\mathbf{r},z)=E\psi(\mathbf{r},z), \tag{6.6}$$

where $m_{//}$ and m_z and are the effective masses, and $\mathbf{r}$ and z are the position coordinates, parallel and perpendicular to the surface. V_{eff} is the potential energy which includes the electrostatic potential due to band bending at the surface, the variation of the conduction band edge with position across the Si/SiO_2 interface, and many-body contributions to the one-electron energy in

terms of the other electrons in the system, i.e., the Hartree, exchange, and correlation terms. Since the potential is only in the z direction, the solution is separable, with free electron motion in the plane parallel to the interface and quantized motion perpendicular, such that the total energy relative to the Si conduction band minima is written as

$$E_{n,\mathbf{k}} = E_n + \frac{\hbar^2 k^2}{2m_{//}}, \tag{6.7}$$

where E_n is the nth subband level, found from solution of the 1D envelope function equation coming from Eq. (6.6).

By the early 1980s, considerable advance had been achieved in the atomic layer growth of semiconductors using techniques such as molecular beam epitaxy (MBE), which allowed control of heterostructure layers within a single atomic layer. Hence, high-quality heterostructure systems could be grown, in which quantum confinement was provided by the band offsets between nearly lattice matched semiconductor systems, such as GaAs and $Al_xGa_{1-x}As$. Figure 6.12 illustrates a *quantum well* structure formed by layering a narrower gap material (material B, e.g., GaAs) by wider bandgap materials (A, e.g., AlGaAs). The bandgap difference between the two materials occurs partly in the valence band and partly in the conduction band. For the $GaAs/Al_xGa_{1-x}As$ system, the band offset in the conduction band has been found empirically to be approximately 65% of the total bandgap difference between the two materials. For other material systems, this ratio is of course different. The system shown in Fig. 6.12 is referred to as a Type I system in that both the valence band and conduction band in the wider gap material form a barrier to carriers (electrons and holes) localized in the quantum well. The solution of the single-band envelope function equation leads to a finite set of independent subbands for

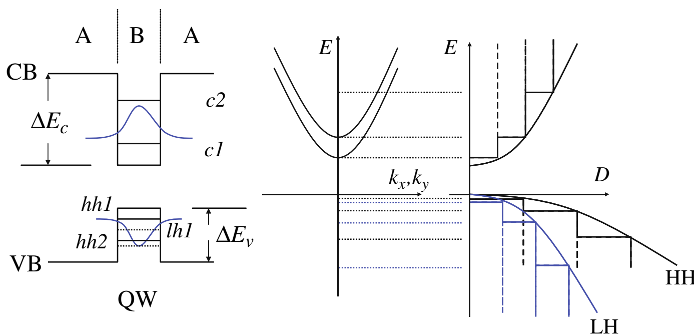


Fig. 6.12 Type I quantum well (QW) structure in a typical III–V compound heterostructure. The left side is the conduction band (CB) and valence band (VB) minima and maxima, respectively, with the corresponding subband structure and density of states for the CB, light hole (LH), and heavy hole (HH) VBs

the CB electrons, and the light hole and heavy hole valence band holes. In reality, for holes, single band theory is insufficient due to strong mixing of the light hole and heavy hole states by the confining potential, and a multi-band or more ab initio approach must be taken to calculate the electronic states [41].

The simple QW system illustrated in Fig. 6.12 corresponds to the case of low doping and carrier concentration in the well. In order to provide free carriers for transport without degrading the carrier mobility due to ionized impurity scattering due to dopants, the concept of *modulation doping* was introduced [42], as illustrated by the inset in Fig. 6.13a. In this scheme, doping is only introduced into the wider bandgap barrier material (e.g., AlGaAs), with a *spacer layer* of undoped barrier material between the doped region and the QW. The free electrons excited to the CB edge of the barrier material from ionized donors there fall into the lower energy states of the well and are confined there. The special separation of the ionized dopants from the free carriers in the well greatly reduces the cross-section for scattering due to the decay of the screened Coulomb potential with distance. Clearly the greater the spacer layer thickness, the less influence scattering due to *remote impurities* is, but at the cost of reduced transfer efficiency from the barrier to the well of free carriers.

Figure 6.13b shows the measured mobility versus temperature in bulk GaAs, and by various groups for modulation doped structures in different years [43],

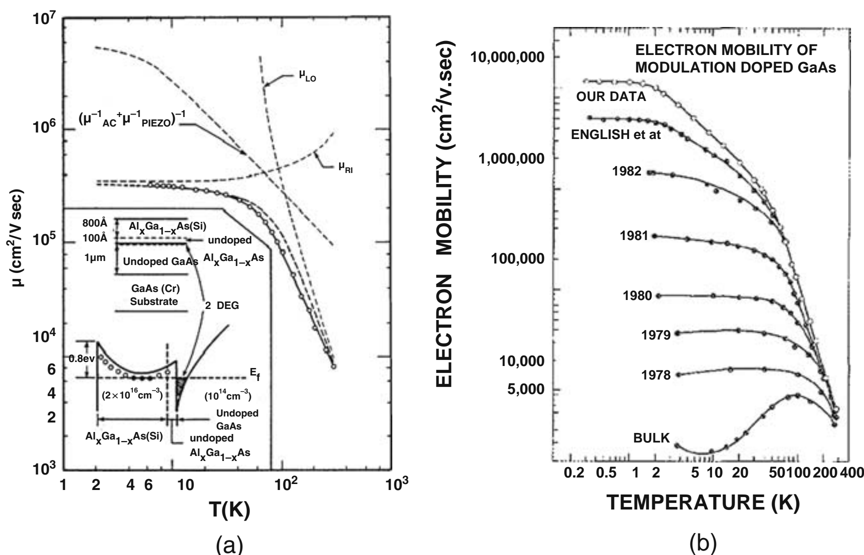


Fig. 6.13 Mobility in a modulation doped AlGaAs/GaAs heterostructure: (**a**) the contribution of individual scattering mechanisms as a function of temperature. Reproduced with permission from Ref. [44], Copyright 1984, American Institute of Physics); (**b**) different results over different years (reproduced with permission from Ref. [43], Copyright 1989, American Institute of Physics

while Fig. 6.13a shows the calculated 2D scattering rates due to individual mechanisms versus temperature, fit to a particular set of data [44]. The different scattering mechanisms shown in Fig. 16.3a correspond to longitudinal optical phonons (LO), acoustic and piezoelectric phonon mode scattering, and remote impurity (RI) scattering. One difference that can be observed between bulk and quasi-2D scattering is that RI scattering basically remains flat with temperature below a certain value, as opposed to decreasing monotonically, which is partly associated with the 2D nature of scattering. The increases with mobility versus time in the right hand figure, reflects the increased purity of as-grown material as well as increasing optimization of the modulation doping itself, as epitaxial growth processes improved with time.

6.3.2.2 Transport in Quasi-1D Systems

Quasi-1D systems may be realized experimentally through top-down fabrication using high-resolution lithographic processes (e.g., electron-beam lithography), to pattern a heterostructure or oxide–semiconductor structure as illustrated in Fig. 6.14a. A modulation doped AlGaAs/GaAs structure, for example, may be etched in various ways to define a nanowire structure as shown, or metallic electrodes may be patterned, and negatively biased to pinch off the 2D electron gas everywhere except between the electrodes, forming a quasi-1D system. Many of the early studies of 1D transport rely on such laterally patterned structures.

More recently, a great deal of interest has been generated by demonstration of directed self-assembly of NWs via in situ epitaxial growth [45, 46], illustrated in Fig. 6.14b. As discussed in the introduction, such semiconductor NWs can be elemental (Si,Ge) or III–V semiconductors, may be controllably doped during growth [31], and high-quality 1D heterojunctions can be achieved. Nanowire FETs, bipolar devices, and complementary inverters have been synthesized

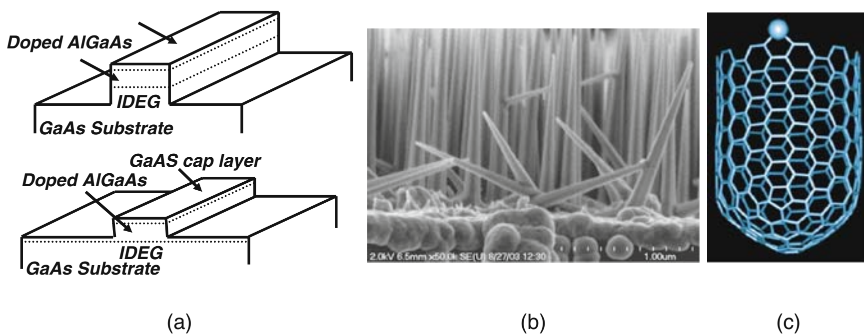


Fig. 6.14 Different experimental realizations of quasi-1D systems. (**a**) Different structures realized through lateral etching or confinement of a 2D quantum well structure. (**b**) A self-assembled Si nanowire structure grown using vapour–liquid–solid epitaxy. (**c**) A carbon nanotube (*See* Color Insert)

using such techniques [24, 47, 48]. The ability to controllably fabricate heterostructure nanowires has led to demonstration of nanoelectronic devices such as resonant tunneling diodes (RTDs) [49] and single electron transistors (SETs) [50]. The scalability of arrays of such nanowires to circuits and architectures has also started to be addressed [51].

Likewise, carbon nanotubes (CNTs) have received considerable attention due to the ability to synthesize NTs with metallic, semiconducting, and insulating behavior, depending primarily on the chirality (i.e., how the graphite sheets forming the structure of the CNT wrap around and join themselves) [34]. Figure 6.14c illustrates a typical CNT structure. Complementary n- and p-channel transistors have been fabricated from CNTs, and basic logic functions demonstrated [52]. The primary difficulty faced today is the directed growth of CNTs with the desired chirality, and positioning on a semiconductor surface, suitable for large-scale production.

Transport in quasi-1D systems such as those discussed above, differs from 2D and bulk in terms of the further reduction in dimensionality, with the corresponding density of states shown in Fig. 6.9b, which is singular at the subband edge. Intrasubband scattering can only have two possible final states, forward or backward along the axis of the wire. The reduced phase space for scattering has often been used as an argument for predicting high mobilities in such systems. However, in lithographically defined systems, the disorder induced by the fabrication process itself usually results in the opposite effect. Self-assembled structures such as the CNTs and semiconductor NWs help avoid process-induced disorder, and in fact very long mean free paths have been observed in CNTs. In addition to the confinement imposed on the electron system in such structures, the vibrational modes of the system are also greatly modified, which in turn affects the transport properties as well.

Transport in Si Nanowires

As an example of transport in a quasi-1D system, we consider the theoretical transport through a top-down fabricated Si nanowire (SiNW) structure [53], as illustrated in Fig. 6.15a, corresponding to the structure originally proposed by Majima et al. [54]. This structure is fabricated from a Si on insulator (SOI) structure, in which a thin SOI layer on a thick buried oxide (BOX) is patterned laterally to form a SiNW as shown. A cutline of the potential perpendicular to the surface through the SiNW is shown in Fig. 6.15b, showing the vertical potential confinement seen by electrons confined to the Si channel region.

The potential and electronic structures are found by solving the coupled 2D Poisson–Schrödinger equations along slices in the AB direction as indicated in Fig. 6.15a. A semi-classical ensemble Monte Carlo simulation was used to simulate particle transport along the wire, assuming the wire is long enough that the diffusive regime of transport is appropriate. Phonon scattering due to bulk acoustic and optical intervalley phonons was included, and surface roughness scattering (SRS) associated with the side walls and the Si–SiO_2 interface

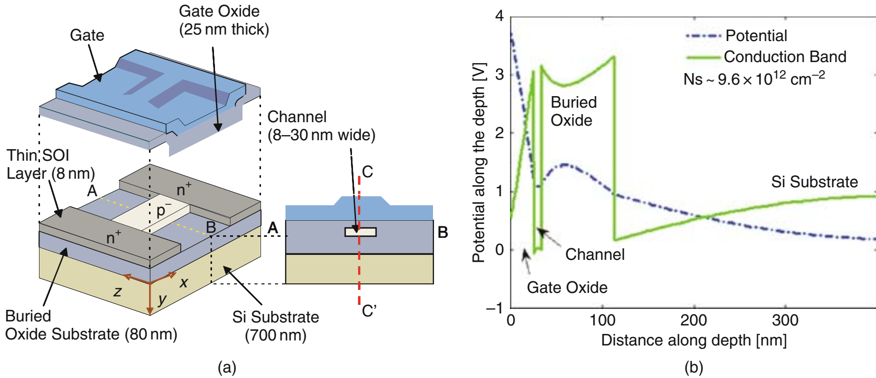


Fig. 6.15 The *left panel* (**a**) shows the schematic of a simulated SiNW on ultrathin SOI. The conduction band profile on the *right side* (**b**) is taken along the red cutline CC from the top panel. The width of the channel is 30 nm [53] (*See* Color Insert)

[40], modified for a quasi-1D system. Full intrasubband and intersubbband scattering was included as well.

Figure 6.16 shows the calculated variation of the mobility with the gate potential as measured by the effective field at the Si–SiO_2 interface, for three different wire widths. For effective fields less than 0.1 MV/cm, scattering is dominated by acoustic and intervalley phonons, and the mobility shows a size-dependent reduction of the mobility with decreasing wire size, due to the effects of confinement. At higher effective fields, the mobility actually increases with decreasing wire width, due to the influence of confinement on surface roughness scattering.

One can understand this mobility enhancement effect at high effective fields, by considering the calculated charge distribution from the coupled 2D Schrödinger–Poisson equation shown in Fig. 6.17. For wider wires,

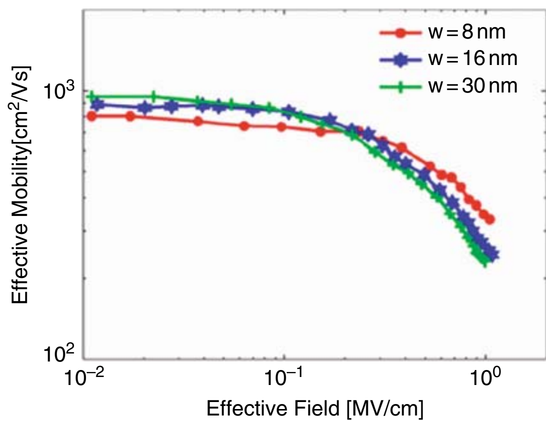


Fig. 6.16 Variation of the field-dependent mobility with varying SiNW width. The wire thickness is kept constant at 8 nm [53] (*See* Color Insert)

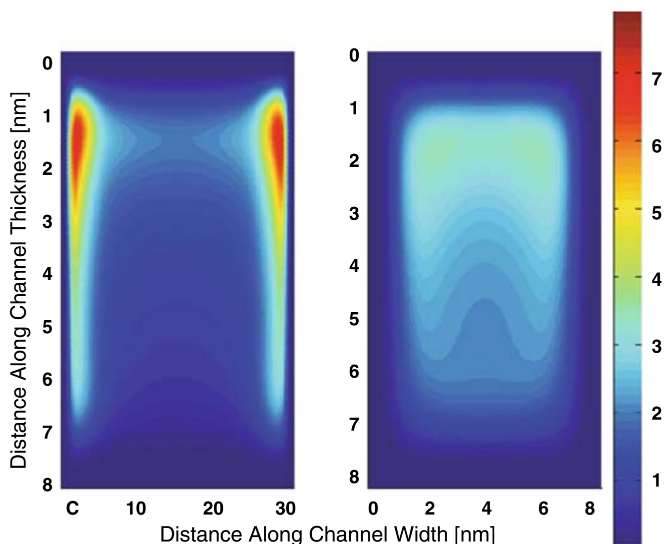


Fig. 6.17 Electron distribution across the nanowire, for the wire width of 30 nm (*left panel*) and 8 nm (*right panel*). In both panels, the transverse field is 1 MV/cm, the wire thickness is 8 nm, and the color scale is in $\times 10^{19}$ cm^{-3} [53] (*See* Color Insert)

electrons are localized close to the Si sidewalls, where surface roughness scattering is effective. However, for narrower wires, there is the onset of volume inversion, the electron wavefunction is localized in the center rather than the edge of the wire, decreasing the effect of roughness scattering and increasing the mobility.

As mentioned earlier, in nanostructures, the phonon dispersion can be affected as well as the electronic structure, leading, e.g., to a modification of scattering and hence transport. Figure 6.18a shows the calculated acoustic phonon dispersion in the presence of confinement using a dielectic continuum model [55]. As can be seen, the simple acoustic dispersion (which goes to zero as the wavevector goes to zero), now shows a number of non-zero branches due to the effect of zone-folding of the dispersion, and the presence of confined phonon modes, similar to an acoustic waveguide. As a result, there are many more channels for scattering, which is reflected in the calculated phonon scattering rates shown in Fig. 6.18b in comparison to the bulk phonon scattering rate. There are many more peaks in the scattering rate due to the contributions of individual modes, superimposed on the quasi-1D scattering rates.

Transport in Self-Assembled Semiconductor Nanowires

Vapor–liquid–solid (VLS) nanowires (NW) may be grown directly on Si substrates using nanoscale liquid metal seeds. A gas-phase precursor, such a silane

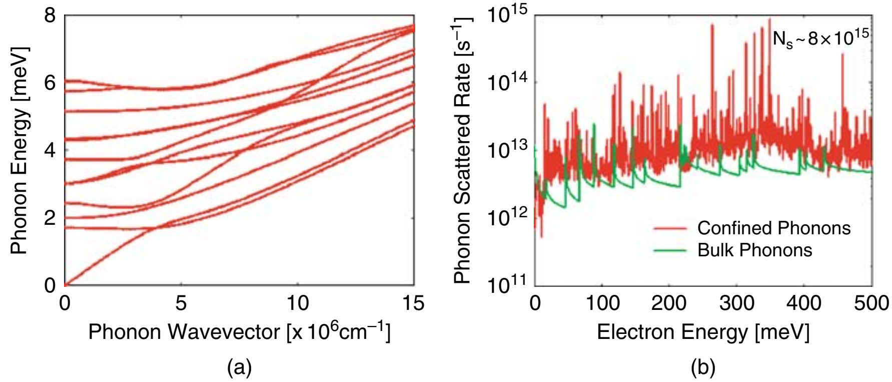


Fig. 6.18 Effect of confinement on the phonon dispersion in a SiNW (**a**), and the corresponding effect on the quasi-1D scattering rate (**b**) (*See* Color Insert)

or germane, transports the material of interest to the seed particle where it dissolves, forming a liquid metal eutectic, as shown in the top panel of Fig. 6.19. The size of this eutectic seed fixes the NW diameter. As the NW crystallizes at the liquid/solid interface, the seed particle 'floats' at its tip. VLS growth produces high-quality, single crystal NWs and heterostructures that are electrically contacted to the substrate. As illustrated in the bottom panel of Fig. 6.19, a variety of structures may be grown, starting with a homogeneous single crystal wire on the left, a vertical heterostructure in the next sketched, followed by the so-called *core-shell* heterostructures consisting of different materials at different radial distances.

Nanowire field effect transistors (FETs) have been synthesized using the above growth techniques by Cui et al. [47] by dispersing the nanowires on a substrate and patterning contacts to two ends of the nanowire. An illustration of a Si nanowire (SiNW) FET structure is shown on the left hand side of Fig. 6.20. SiNWs are dispersed on an oxidized conducting semiconductor substrate, where the substrate itself is used as a gate as shown. TiAu contacts

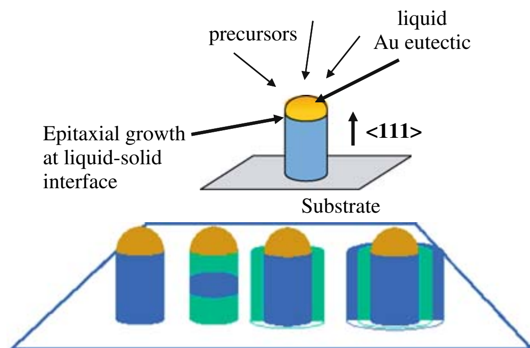


Fig. 6.19 Schematic of growth of a semiconductor nanowires using vapor–liquid– solid (VLS) phase growth. The *bottom panel* illustrates several different heterostructures realizable using this technique (*See* Color Insert)

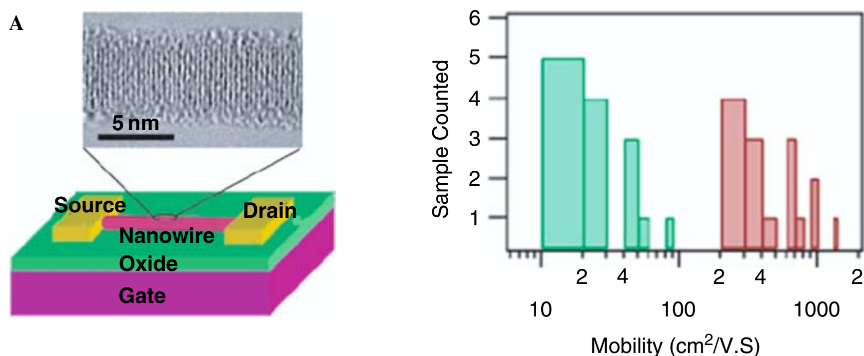


Fig. 6.20 Si Nanowire field effect transistor structure. The *left panel* shows a schematic and electron micrograph of the transistor structure. The *right panel* shows the measured mobility before (left side) and after (right side) surface modification. Reprinted with permission from Ref [47], Copyright 2003 (*See* Color Insert)

are deposited for the source and drain ohmic contacts. A high-resolution transmission electron microscope (HRTEM) micrograph of an approximately 5 nm diameter wire is shown as well. As can be seen, the wire cross-section shows a high degree of crystallinity, with an amorphous surface layer evident. A certain degree of surface roughness is present as well, although significantly less than that found from top-down patterning, particularly at such narrow dimensions.

Transport in semiconductor nanowires is quasi-1D, depending on the diameter of the wire. For Si wires greater than 20 nm, it is expected that transport will be more bulk-like due to many occupied subbands, whereas smaller diameters should exhibit significant quantization of motion. Due to the small diameter in self-assembled structures, transport is very sensitive to the structure of the surface, the degree of roughness, the presence of traps or interface charges due to dangling bonds, etc. Contacts and contact resistance are another issue which has to be separated from transport in the nanowire itself. Cui et al. studied a number of different passivation techniques to improve the mobiliy; the left side of Fig. 6.20 shows the measured mobility before (left set of data) and after modification with 4-nitrophenyl octadecanoate. Clearly an enormous improvement in mobility is observed, suggestive of the strong role played by the surface in transport.

Transport in Carbon Nanotubes

Single-walled (SW) carbon nanotubes (CNTs) are a tubular form of carbon with diameters as small as 1 nm and lengths of a few nanometers to microns [34, 56]. Typically, CNTs are grown by chemical vapor deposition (CVD), laser ablation, or arc discharge processes. A CNT is configurationally equivalent to a 2D graphene sheet rolled into a tube, as shown in Fig. 6.21. In rolling the tube,

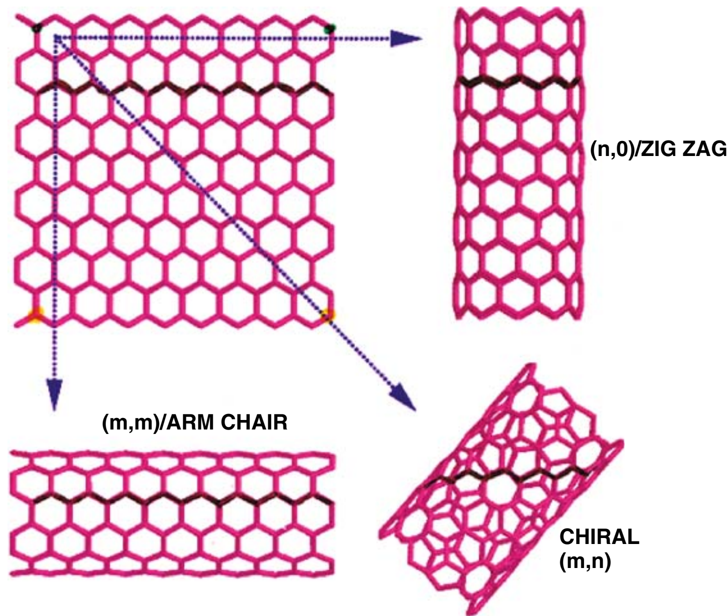


Fig. 6.21 Molecular structure of a single-wall CNT, formed by rolling a sheet of graphene, illustrating different chiralities (*See* Color Insert)

the direction along which the tube is rolled to form a closed structure defines its *chirality*, defined by a pair of indices (n,m) called the chiral vector. The integers n and m represent the number of unit cells along the 2D hexagonal forming the graphene lattice. If $m = 0$, the nanotubes are referred to as 'zigzag', while $n = m$, corresponds to the so-called 'armchair' structure.

The electronic and transport properties are strongly dependent on the chirality and diameter of the CNT. If for a given chiral vector, $n–m$ is a multiple of 3, then the nanotube is metallic, which includes the armchair ($n = m$) structure shown in Fig. 6.21. The conductivity of metallic CNTs is quite high due to the high mobility and density. Other chiral structures not satisfying this equality behave as semiconductors. The bandgap is roughly proportional to 1/diameter, with $E_g = 0.5$ eV for a diameter of 1.4 nm. Besides their unique electronic properties, CNTs exhibit extraordinary mechanical properties, with a Young's modulus over 1 TPa, as stiff as diamond, and tensile strength $\sim$200 GPa.

Due to their remarkable electronic and mechanical properties, CNTs are currently being developed for a number of applications including interconnects, CNT-based molecular electronics, AFM-based imaging, nanomanipulation, nanotube sensors for force, pressure, and chemical, nanotube biosensors, molecular motors, nanoelectromehanical systems (NEMS), hydrogen and lithium storage, and field emitters for instrumentation including flat panel displays.

In terms of transport, measurements have demonstrated very high mobilities and nearly ballistic transport [57, 58]. In this context, a diffusive picture of transport in CNTs is not appropriate, rather a treatment in terms of quantum fluxes as discussed in Section 6.4. However, dopants and defects can lead to scattering. CNTs are inherently p-type, but by annealing in vacuum or doping with electropositive element (e.g., K), they can be doped n-type. Electron–electron can contribute to scattering. While normally conserving the net momentum of the two particles, and hence not relaxing the net momentum, Umkapp processes are possible within the reduced zone of the CNT bandstructure, which do lead to a net backscattering [59]. There are various other mechanisms that limit transport, particularly at high fields when electrons are accelerated above the threshold for various types of phonon scattering. In particular, because of the unique hollow structures of CNTs, there are torsional modes of vibration, similar to molecular chains, such as twistons, which are essentially long wavelength acoustic phonon modes [60]. Optical modes associated with the in-plane modes of graphene and zone boundary phonons coupling different Fermi wavevectors are also believed to be important [61]. Studies of high field transport in metallic SW CNTs with low contact resistance show that the saturation of current at high bias is associated with optical and zone boundary phonon emission [61].

Summary of Diffusive Transport in Nanowires and Nanotubes

Summarizing this section on transport in nanowires and nanotubes, some of the observations that one can make are the following:

- Predictions that transport should be improved in nanowires due to reduction of phase space for scattering, and reduced coupling to phonons (bottleneck effects).
- Measured mobilities in lithographically defined nanowires are less than the bulk due to process-induced roughness and other inhomogeneities.
- Semiconductor nanowires show higher effective mobilities than etched structures, but limits occur due to surface states and surface morphology, as well as contact issues.
- CNTs show high conductivity, effective mean free paths of several microns at room temperature inferred. Difficulty in extracting CNT resistance from contact resistance.

6.4 Transmission and Transport in Nanoscale Systems

As discussed earlier, when length scales become sufficiently short that the quantum mechanical phase becomes important, the nature of transport changes dramatically from one derived from a semi-classical, particle-based picture of scattering to one where transport is determined by the quantum mechanical flux through the system, and the reflection and transmission of carriers injected into the system. We

start first with the case of 1D quantum transport through potential barriers, followed by a discussion of transport in the technologically important resonant tunneling diode in Section 6.4.1. We then proceed to discuss transport in quantum point contacts and quantum waveguide structures within the context of the Landauer–Büttiker model for conductance in Sections 6.4.2 and 6.4.3.

6.4.1 Vertical Transport Through Heterostructures

As a prototype system to consider in terms of coherent transport, consider the tunneling of an electron through a planar barrier structure as shown in Fig. 6.22. The applied bias separates the Fermi energies on the left and right by an amount eV. The Hamiltonian on either side of the barrier is assumed separable into perpendicular (z-direction) and transverse components. If we choose the zero-reference of the potential energy in the system to be the conduction band minimum on the left, $E_{\mathrm{c,l}} = 0$, the energy of a particle before and after tunneling may be written as

$$E = E_z + E_t = \frac{\hbar^2 k_{z,\mathrm{l}}^2}{2m^*} + \frac{\hbar^2 k_{t,\mathrm{l}}^2}{2m^*} \tag{6.8}$$

on the left side, and

$$E = E_z + E_t = \frac{\hbar^2 k_{z,\mathrm{r}}^2}{2m^*} + \frac{\hbar^2 k_{t,\mathrm{r}}^2}{2m^*} + E_{\mathrm{c,r}} \tag{6.9}$$

on the right side, where $E_{\mathrm{c,r}}$ is the conduction band minimum on the right side and k_z and k_t are the longitudinal and transverse components of the wavevector relative to the barrier. Assuming the transverse momentum is conserved during

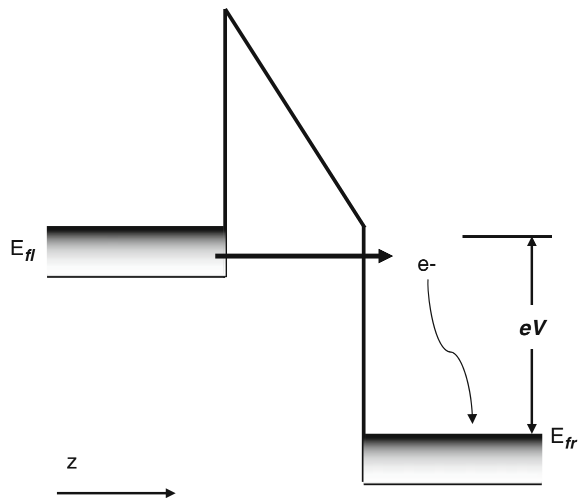


Fig. 6.22 Band diagram for a tunnel barrier under bias, illustrating charge flow

the tunneling process, $k_{t,\mathrm{l}} = k_{t,\mathrm{r}}$, and the transverse energy $E_{t,\mathrm{l}} = E_{t,\mathrm{r}}$ is the same on both sides for the tunneling electron. Therefore, the z-component of the energy on the left and right sides of the barrier is

$$E_z = \frac{\hbar^2 k_{z,\mathrm{l}}^2}{2m^*} = \frac{\hbar^2 k_{z,\mathrm{r}}^2}{2m^*} + E_{\mathrm{c,r}}. \tag{6.10}$$

The incident current density from the left may be written in terms of the flux of carriers arising from an infinitesimal volume of momentum space d**k**, around **k**,

$$j_l = -e\rho(\mathbf{k}_\mathrm{l}) f_\mathrm{l}(\mathbf{k}_\mathrm{l}) v_z(\mathbf{k}_\mathrm{l}) \mathrm{d}\mathbf{k}_l, \;\; \rho(\mathbf{k}_\mathrm{l}) = \frac{2}{(2\pi)^3}, \tag{6.11}$$

where f_l is the distribution function on the left side of the barrier, $\rho(\mathbf{k})$ is the density of states in **k**-space, and the velocity perpendicular to the barrier from the left is (assuming parabolic bands)

$$\nu_z(\mathbf{k}_\mathrm{l}) = \frac{1}{\hbar}\frac{\partial E(\mathbf{k}_\mathrm{l})}{\partial k_{z,\mathrm{l}}} = \frac{\hbar k_{z,\mathrm{l}}}{m^*}. \tag{6.12}$$

The transmitted current density from the left to right is simply Eq. (6.4) weighted by the transmission coefficient

$$j_l = -\frac{2e\hbar}{(2\pi)^3 m^*} T(\mathbf{k}_{z,\mathrm{l}}) f_\mathrm{l}(\mathbf{k}_t, k_{z,\mathrm{l}}) k_{z,\mathrm{l}} \mathrm{d}k_{z,\mathrm{l}} \mathrm{d}\mathbf{k}_t, \tag{6.13}$$

where $T(k_{z,\mathrm{l}})$ is the transmission coefficient from the left. Similarly, the transmitted current from the right to left is

$$j_\mathrm{r} = -\frac{2e\hbar}{(2\pi)^3 m^*} T(\mathbf{k}_{z,\mathrm{r}}) f_\mathrm{r}(\mathbf{k}_t, k_{z,\mathrm{r}}) k_{z,\mathrm{r}} \mathrm{d}k_{z,\mathrm{r}} \mathrm{d}\mathbf{k}_t. \tag{6.14}$$

Integrating over all **k**, and assuming that the distribution functions in the contacts on the left and right are given by their equilibrium Fermi–Dirac distributions corresponding to the Fermi energies in the respective regions, one can integrate over the transverse energies analytically to obtain the so-called *Tsu-Esakt formula*, where the particular form was popularized [62] in connection with resonant tunneling diodes (discussed below):

$$\begin{aligned} J_T &= \int \mathrm{d}k_z \mathrm{d}\mathbf{k}(j_\mathrm{l} - j_\mathrm{r}) \\ &= \frac{em^* k_\mathrm{B} T}{2\pi^2 \hbar^3} \int_0^\infty \mathrm{d}E_z T(E_z) \ln\left[\frac{1 + \exp[(E_{F,\mathrm{l}} - E_z)/k_\mathrm{B}T]}{1 + \exp[(E_{F,\mathrm{l}} - eV - E_z)/k_\mathrm{B}T]}\right]. \end{aligned} \tag{6.15}$$

The derivation of this equation invokes several assumptions common to the general description of coherent transport in nanostructures; first, we can describe the current in terms of the net difference in flux of transmitted carriers through the structure; second, we have 'ideal' contacts or reservoirs (as was shown schematically earlier in Fig. 6.5) that are near equilibrium (since the reservoir is assumed large) and inject carriers into the system according to some prescribed or known distribution function.

A particularly interesting and technologically important planar barrier structure is the double barrier *resonant tunneling diode* (RTD, which was predicted theoretically by Tsu and Esaki [62, 63] at IBM, and demonstrated definitely by Sollner et al. [64]. The conduction band diagram of a generic RTD structure is shown in Fig. 6.23 under different bias conditions, along with the associated current–voltage (*I*–*V*) characteristics at each point. The device is a two-terminal structure consisting of two narrow barriers (e.g., $Al_xGa_{1-x}As$) with heavily doped emitter and collector materials (e.g., GaAs) and a narrow well of narrower gap material separating the two barriers. Resonant tunneling occurs when an electron incident on the left is coincident in energy with the *quasi*-*bound state* formed in the well, resulting in a sharp maxima in the transmission coefficient through a structure. The thickness of the barriers (approximately 1–5 nm) is sufficiently thin that tunneling through the barriers is significant. Depending on the well width and barrier heights, there may exist several such *quasi*-*bound states* in the system. As shown in this figure, with a positive bias applied to the right contact relative to the left, the Fermi energy on the left is pulled through the resonant level.

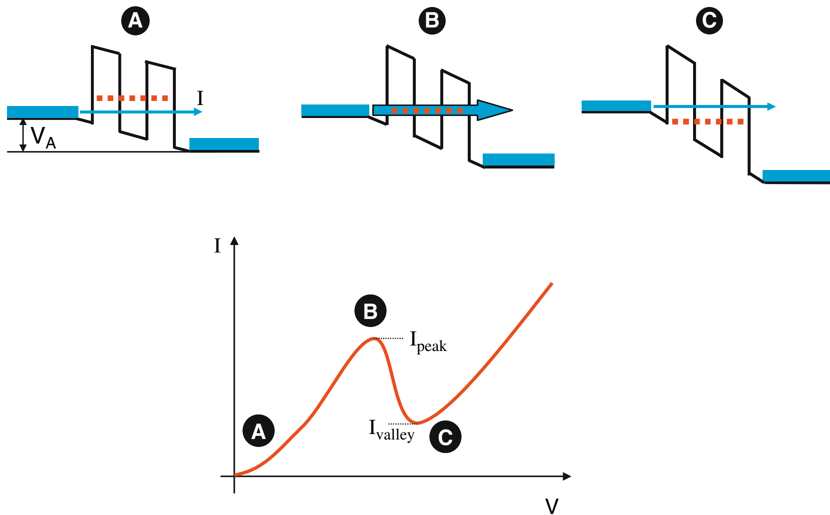


Fig. 6.23 *Upper panel* represents the band diagram of a resonant tunneling diode at various bias points on and off resonance, while the *bottom panel* is the corresponding current

As the Fermi energy passes through the resonant state, a large current flows due to the increased transmission from left to right. At the same time, the backflow of carriers from right to left is suppressed as electrons at the Fermi energy on the right see only a large potential barrier. Further bias pulls the bottom of the conduction band on the left side through the resonant energy, which cuts off the supply of electrons available at the resonant energy for tunneling. The result is a marked decrease in current with increasing voltage, giving rise to a region of negative differential resistance (NDR) as shown schematically by the *I–V* characteristics in Fig. 6.23. A figure of merit for the performance of an RTD is the peak to valley ratio (PVR), which is the ratio of the peak current to the valley current shown below. For AlAs/GaAs/AlAs RTD structures, ratios of 4:1 or more may be realized at room temperature, and much larger at lower temperature since thermal processes tend to increase the off-resonant portion of the current and decrease the resonant portion as discussed below.

The purely coherent picture of transport in RTDs above is of course an idealization, and one of the fundamental issues of transport in nanostructure systems, that is the loss of coherence through interaction with the environment and dissipative processes in the structure itself. Figure 6.24 shows a more realistic band diagram of the RTD under bias illustrating various possible contributions to the current. A process of sequential tunneling is possible [65], in which an electron is injected above the resonant energy, and then relaxes down to the ground state of the well, before tunneling out into

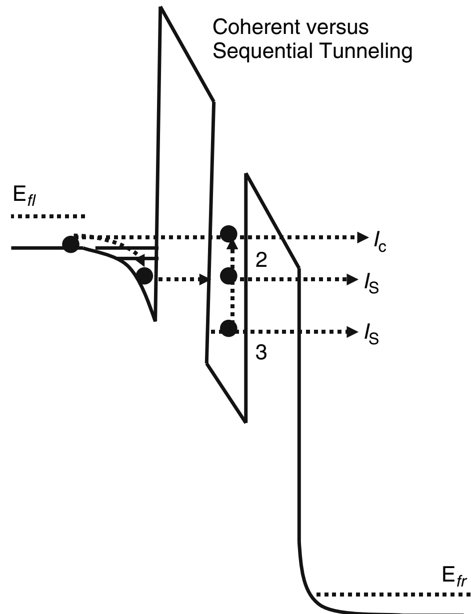


Fig. 6.24 Band diagram of the a resonant tunneling diode under bias showing various contributions to the current, both coherent and sequential

the collector. This process may actually give DC I–V characteristics similar to those for purely coherent tunneling. However, inelastic tunneling processes are possible as well, where electrons may emit or absorb a phonon to complete the tunneling process through the double barrier structure. Elastic scattering processes such as impurities or interface roughness also relax the assumption of parallel momentum conservation, leading to a broadening of the resonance. Thermoionic emission over the top of the barrier is possible as well. All these contributions lead to increased non-resonant current and decreased on-resonant current, leading to a degradation of the peak to valley ratio.

6.4.2 Quantized Conductance

We now consider the general barrier problem of phase coherent transport through a 1D conductor as shown in Fig. 6.25, which, for example, may correspond to the quasi-1D systems discussed in Section 6.3.2.2 in the limit that inelastic (phase breaking) scattering processes are negligible except in the contacts . Ideal (i.e., no scattering) conducting leads connect the scattering region to reservoirs on the left and right characterized by quasi-Fermi energies μ_1 and μ_2, respectively, corresponding to the electron densities there.

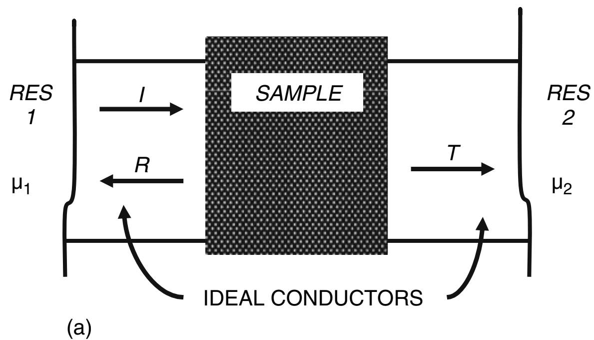


Fig. 6.25 Schematic illustration of the conductance of a 1D system. (**a**) Conceptualization of ideal 1D conductors connecting the 'sample' to infinite reservoirs. (**b**) Redistribution of charge injected by the reservoirs due to scattering from the sample, resulting in new Fermi energies μ_A and μ_B on the left and right sides, respectively

These reservoirs or contacts are assumed to randomize the phase of the injected and absorbed electrons through inelastic processes such that there is no phase relation between particles. As for the planar barrier structures in the previous section, the current injected from the left and right may be written as an integral over the flux in 1D:

$$I = \frac{2e}{2\pi}\left[\int_0^\infty \mathrm{d}k\nu(k)f_1(k)T(E) - \int_0^\infty \mathrm{d}k'\nu(k')f_2(k')T(E')\right], \tag{6.16}$$

where the prefactor corresponds to the 1D density of states in k-space times e, $v(k)$ is the velocity, $T(E)$ is the transmission coefficient, and f_1 and f_2 are the reservoir distribution functions characterized by their respective Fermi energies introduced above. The integrations are only over positive k and k' relative to the direction of the injected charge. If we now assume low temperatures, electrons are injected up to an energy μ_1, into the left lead, and injected up to μ_2 into the right one. Converting to integrals over energy, the current becomes

$$I = \frac{e}{\pi}\left[\int_0^{\mu_1} \mathrm{d}E\left(\frac{\mathrm{d}k}{\mathrm{d}E}\right)\nu(k)f_1(k)T(E) - \int_0^{\mu_2} \mathrm{d}E\left(\frac{\mathrm{d}k'}{\mathrm{d}E}\right)\nu(k')f_2(k')T(E)\right]. \tag{6.17}$$

Since the electron velocity itself is defined as the derivative of E with respect to k (group velocity), this equation reduces to

$$I = \frac{e}{\pi\hbar}\int_{\mu_2}^{\mu_1} \mathrm{d}ET(E) = \frac{e}{h}T(\mu_1 - \mu_2) = \frac{e^2}{h}TV, \tag{6.18}$$

where it has been assumed that T is a weak function of energy between the Fermi energies on the right and left, which is asymptotically true in the linear response regime where the two converge. Defining the conductance as I/V, the so-called two-terminal *Landauer formula* results in [66, 67]

$$G = \frac{2e^2}{h}T, \tag{6.19}$$

which states that the conductance of a pure 1D system is simply a universal constant, $G_o = 2e^2/h$ (or the inverse of the so-called fundamental resistance, equal to 12.9 kΩ) times the transmission coefficient through the system. Note that the Fermi levels in the leads are different from the reservoir Fermi energies due to the buildup and depletion of charge due to scattering from sample, giving rise to the so-called *Landauer resistivity dipole*. If in fact one could measure the potential drop across the sample itself, then Eq. (6.19) would be modified by a factor of $R = (1-T)$ in the denominator, which diverges as R goes to 0. However,

in the limit of unity transmission, the two-terminal conductance measured between the two reservoirs does not diverge, but reaches a finite value, which represents the *contact* resistance to inject charge in and out of the 1D system.By extending the above discussion to N occupied subbands (channels), the general two-terminal *multi-channel* Landauer–Büttiker [68] formula results in

$$G = \frac{2e^2}{h}\sum_{n=1}^{N} T_n, \tag{6.20}$$

where T_n is the total transmission through channel n to all other channels. In the limit of T_n going to unity, the conductance simply becomes NG_o.

The first demonstrations of this conductance quantization were provided by studies of *split-gate* quantum point contacts, which were implemented in the high-mobility 2D electron gas of GaAs/AlGaAs heterojunctions [7, 8], as shown in Fig. 6.26. Here, metal gates are patterned on the top of the heterojunction, and are used to define the constriction in the 2D electron gas underneath, by the application of a negative depleting voltage. The size of the constriction that forms in the electron gas is determined by the range of the fringing fields that develop around the gate edges, and may be tuned continuously in experiment by variation of the gate voltage. As the width of the constriction is reduced in this manner, successive subbands are depopulated and the conductance of the point contact decreases in integer steps of $2e^2/h$.

Figure 6.27 shows the measured conductance on an exceptionally well-resolved sample [69]. The split-gate structure in the inset is used to pinch off and open the channel. The low temperature conductance shows a series of plateaus with conductance given by NG_o. As the sample temperature increases, the distribution functions appearing in Eqs. (6.16) and (6.17) broaden, and the corresponding plateau is washed out. Similarly, if tunneling through the QPC is too strong (i.e., the constriction too narrow), the conductance plateaus are less well resolved.

The presence of disorder has a detrimental effect on the observation of quantized conductance, due to the effect of backscattering, which degrades

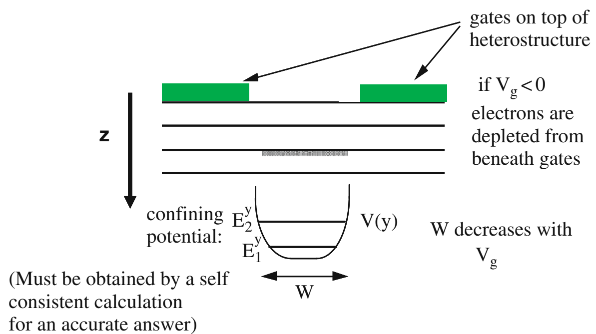


Fig. 6.26 Schematic of a split-gate quantum waveguide or quantum point contact structure for measuring ballistic conductance in quasi-1D systems

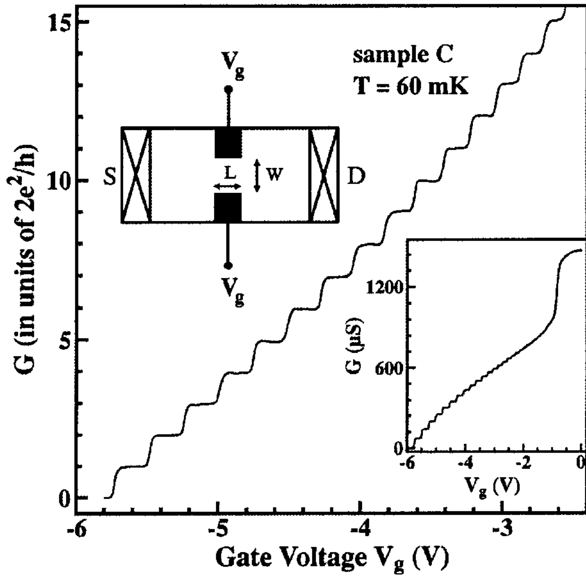


Fig. 6.27 Conductance quantization in the quantum point contact. The upper inset shows a schematic of the device geometry. Reprinted with the permission from Ref. [69], Copyright 1998 by the American Physical Society

the transmission below unity, hence degrading the conductance. Figure 6.28 shows a comparison of two different quantum point contacts grown fabricated on the same material, but with different lengths [70]. For the short length QPC shown in Fig. 6.28a, well-resolved conductance plateaus are observed. For longer constrictions (>500 nm), the conductance plateaus are washed out, and resonant-like structures appear which are related to unique signature of individual scatters associated with impurities and roughness of the QPC itself. As the constriction becomes longer, the probability that an electron is transmitted

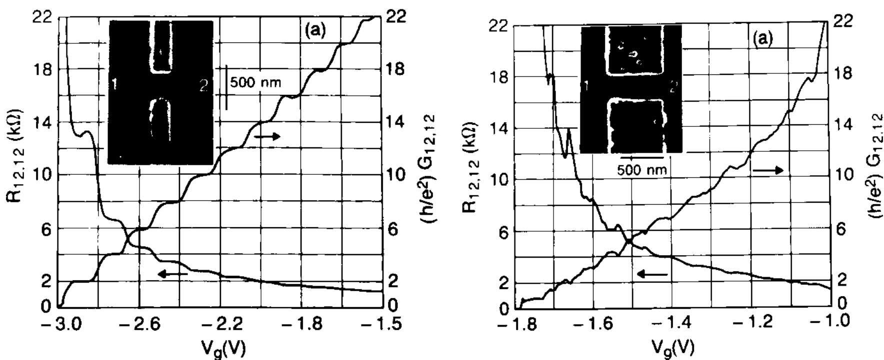


Fig. 6.28 Measured conductance in a quantum point contact structure comparing narrow (*left*) and wide (*right*) constrictions. Reprinted with permission from Timp et al. in Nanostructure Physics and Fabrication, edited by W. P. Kirk and M. Reed., Academic Press, New York, 1989, pp. 331–346. Copyright 1989, Academic Press

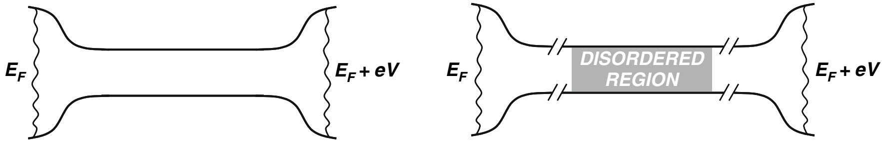


Fig. 6.29 Illustration of an ideal quantum waveguide structure (*left*), and one with a disorder region. Reprinted with permission from R. Akis, private communication)

ballistically without scattering with an unintentional defect is decreased, leading to the transition from ballistic to diffusive behavior.

To illustrate this transition from ballistic to diffusive behavior mathematically, consider the ideal and disordered 1D conductors shown in Fig. 6.29. The resistance, R, through the disordered system, can be approximated by assuming each conducting channel has the same transmission coefficient, T_{pc}:

$$R = \frac{1}{G} = \frac{h}{2e^2}\frac{1}{NT_{\mathrm{pc}}} = \frac{h}{2e^2}\frac{1}{N} + \frac{h}{2e^2}\frac{1}{N}\left[\frac{1-T_{\mathrm{pc}}}{T_{\mathrm{pc}}}\right] = \frac{1}{G_{\mathrm{C}}} + \frac{1}{G_{\mathrm{D}}}, \qquad (6.21)$$

which can be factored into one term which is the pure ballistic conductance, G_{c}, and a second which is the conductance determined by the transmission and reflection through the disordered region itself. As T_{pc} goes to zero (meaning strong disorder), the second term dominates, and the conductance is due entirely to G_{D}. As T_{pc} approaches unity, the disorder term vanishes, and we are left with the quantized conductance associated with coherent transport.

6.4.3 Quantum Waveguides

The coherent transport of electrons through quasi-1D channels is quite analogous to the transmission of electromagnetic radiation through waveguides in terms of transmission and reflection at different wavelengths. Hence one may think of more complicated structures that behave as *quantum waveguides*. As an example of a quantum waveguide, consider the device structure shown in Fig. 6.30, which consists of split-gate structure which has been patterned to create double-bend discontinuity. In Fig. 6.30a, we show a micrograph of the structure, and the measured variation of the waveguide conductance as a function of the gate voltage [71]. Rather than exhibiting quantized plateaus, there appear to be a series of resonances below the fundamental conductance, superimposed on a background conductance. In Fig. 6.30b, the computed [73] variation of the conductance using a scattering matrix approach as function of the width of an ideal, hard-walled waveguide structure is shown. The conductance here is calculated by first calculating the transmission coefficients as a function of energy, $T_n(E)$, from scattering theory, and then applying the Landauer–Büttiker formula at the appropriate Fermi energy. As can be seen, strong resonances are calculated due to the standing wave interference patterns

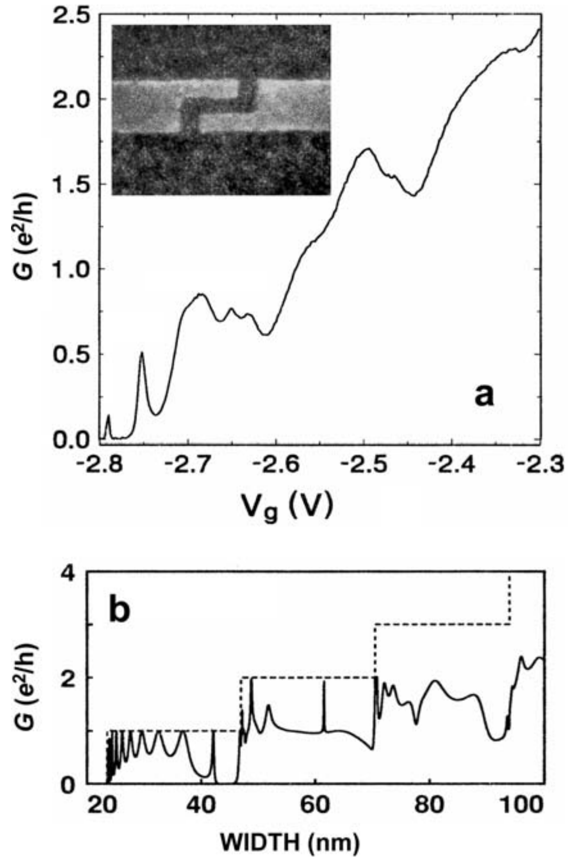


Fig. 6.30 (**a**) Experimentally measured conductance characteristic of an electron waveguide featuring a double-bend structure. The device structure is shown in the inset, in which the lithographic width of the waveguide is 100 nm. (**b**) Numerically calculated conductance variation for a hard-walled waveguide structure

established by the cavity in the double bend. The computed resonances are much sharper than those seen in experiment, but this difference can be attributed to the finite temperature at which the latter is performed, and the soft-walled nature of the confining profile in real devices [73].

A technological important example of a 1D electron waveguides is the three-terminal ballistic junction (TBJ), or Y-branch switch [72, 73, 74], shown in the inset of Fig. 6.31. As a switch, electrons incident on the junction through the center terminal, C, may be deflected either into the right or left branch by application of an electric field. Since switching occurs simply by deflecting a current either left or right, considerable interest has arisen in the potential application of the Y-branch as a fast switch with low power consumption. According to recent theoretical studies [75, 76], the Y-branch may also be used to generate rectifier and transistor behavior, to provide a means of second-harmonic generation and to serve as an oscillator in the THz regime.

An interesting demonstration of the non-classical, coherent transport behavior of TBJs is observed when the voltage or current in the center terminal is

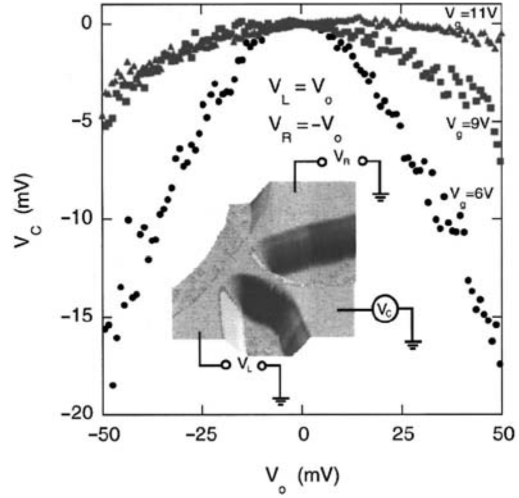


Fig. 6.31 Measured center potential in a three-terminal ballistic junction (TBJ) as a function of the push–pull potential, $V_o = V_L = -V_R$, for various top gate bias voltages. The inset shows a micrograph and schematic of the bias configuration of the structure. Reproduced with permission from Ref. [79], Copyright 2001, American Institute of Physics [81]

measured in the non-linear in a push–pull sort of bias configuration applied to the left and right contacts, $V_R = -V_L = V_o$ [76, 77]. For perfectly symmetric waveguides, the classical prediction is that the center potential should be zero. However, in the quantum mechanical picture, carriers are injected from the left and right terminals into the center terminal, resulting in the development of a negative potential there. Figure 6.31 shows the measured center potential, V_C, as a function of the push–pull potential, V_o. A top gold gate deposited over a dielectric covering the structure is used to vary the Fermi potential, with less positive potentials corresponding to lower Fermi energies in the structure. As the Fermi energy is reduced, one can observe that the center potential becomes increasingly negative with increasing magnitude of V_o. Using an extension of the Landauer–Büttiker formula to multi-terminals, Xu [77] has shown that the center potential may be written as

$$V_c = -\frac{1}{2}\alpha V_o^2 + O(V^4), \quad \alpha = e\frac{\partial G_C(\mu.T)/\partial\mu}{G_C(\mu.T)}, \tag{6.22}$$

where $G_C = G_o[T_{CL}(\mu, T) + T_{CR}(\mu, T)]$ is the conductance seen from the center terminal. Basically, as the Fermi energy is reduced with decreasing top gate potential, the center conductance, G_C, is reduced, increasing α, and hence the enhanced downward curvature as observed in Fig. 6.31 for gate biases of 9 and 6 V. Such downward parabolic variation of the center voltage with push–pull voltage has been observed in a number of experiments [77, 78, 79]. The only assumption made in deriving Eq. (6.22) is that the center cavity should be ballistic, and therefore clear evidence of the non-classical voltage variation has been reported at room temperature [79, 81]. Ballistic TBJs have also been predicted [77, 78] to show rectification and basic transistor action,

second-harmonic generation, and logic operation. In particular, with biases applied to left and right reservoirs, the output voltage of the center waveguide will only be positive when a positive voltage is applied to *both* the left and right reservoirs, indicating that the Y-branch may be used as a compact AND gate. Based on such concepts, there has been considerable interest in the development of novel circuit architectures, based upon the properties of the TBJs [80].

6.5 Single Electron Tunneling

In the previous sections, we discussed phenomena at the nanoscale associated with quantum mechanical effects such as phase coherent transport and size quantization. Another consideration in ultrasmall structures is the granularity of charge itself in terms of the finite number and charge of electrons. Single electron tunneling is a term used to describe the correlated tunneling of electrons one at a time, due to the effect itself on the energy of the system of the motion of a single charge into and out of a nanostructure system. Such single electron effects have been the basis for a number of device and architectural proposals and demonstrations such as single electron memories, single electron transistors, quantum cellular automata, as well as many others. The interested reader is referred to detailed reviews of single electron phenomenon and devices (see, for example, [4, 81, 82]). In the following, we briefly review single electron phenomena in semiconductor systems including some recent results in self-assembled systems.

6.5.1 *Single Electron Phenomena*

Single electron phenomena can be understood classically in terms of the capacitance, C, which relates the charge to the potential difference between two conductors, and the corresponding electrostatic energy, E, stored in the two-conductor system:

$$Q = CV, \qquad E = \frac{Q^2}{2C}. \tag{6.23}$$

Capacitance is reduced as physical dimensions decrease, and for sufficiently small capacitance, a change in charge, $Q = e$, corresponding to the transfer of a single electron, can result in a sizeable change in the electrostatic energy. As an example, the capacitance of a conducting sphere of radius a above a ground plane is approximately $C = 4\pi\varepsilon\, a$. For a 5 nm radius nanocluster (for example, Au forms stable cluster shells of even smaller dimensions), the change of energy associated with removing an electron from an initially charge neutral cluster corresponds to approximately 145 meV, which is much larger than the thermal energy, even at room temperature.

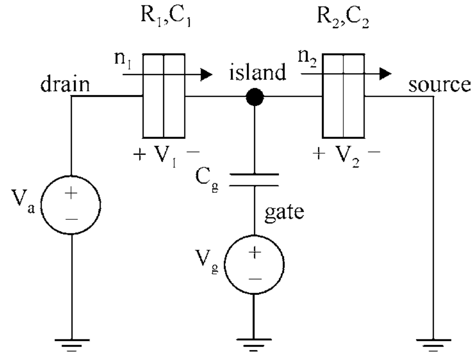


Fig. 6.32 Equivalent circuit for a single electron transistor (SET)

The basic building block of single electron devices and circuits is the tunnel junction, illustrated by the circuit elements shown in Fig. 6.32, which illustrates the schematic of a single electron tunneling transistor (SET). A tunnel junction is characterized by its capacitance, C, and tunnel resistance, R, the latter of which corresponds in the usual generic way to the height and width of the potential barrier between electrodes. Tunnel junctions are actually representative of a broad range of permeable device technologies including ultrasmall metal–oxide–metal junctions (oxidized Al for example), the quantum point contact structure discussed in Section 6.4.2, sidewall constrictions in an etched Si on insulator or GaAs/AlGaAs structure, and even the contacts to carbon nanotubes as discussed later.

The SET transistor, first realized experimentally by Fulton and Dolan [83] and Kuz'min and Likharev [84], consists of a pair of tunnel junction separated by an island with an applied source–drain bias as shown in Fig. 6.32. The island itself (which represents an isolated conducting region) is capacitively coupled through C_g to a gate bias, making a three-terminal structure. To understand the behavior, we consider the energy change when electrons tunnel back and forth across the two tunnel junctions. Here we let n_1 be the net number of electrons that tunneled through the first junction onto the island, n_2 the number of electrons that tunneled through the second junction exiting the island, and $n = n_1 - n_2$ the net number of excess electrons on the island.

With a bias voltage applied across the two junctions, the charges on the junctions and island can be written as

$$Q_1 = C_1 V_1, \quad Q_2 = C_2 V_2 \;\Rightarrow$$
$$Q = Q_2 - Q_1 + Q_0 + C_g(V_g - V_2) = -ne + Q_0 + C_g(V_g - V_2), \tag{6.24}$$

where Q is the net charge on the island, Q_0 is the background charge induced by stray capacitances associated with material imperfections and fabrication-induced defects, and the effect of the gate electrode is to contribute an additional controllable polarization charge on the island.

In terms of the applied bias, $V_a = V_1 + V_2$, we can rewrite the junction potentials as

$$V_1 = \frac{(C_2 + C_g)V_a - C_g V_g + ne - Q_0}{C_\Sigma}, \quad V_2 = \frac{C_1 V_a + C_g V_g - ne - Q_0}{C_\Sigma}, \tag{6.25}$$

where $C_\Sigma = C_1 + C_2 + C_g$. The electrostatic energy stored in the two junctions is

$$E_c = \frac{C_g C_1 (V_a - V_g)^2 + C_1 C_2 V_a^2 + C_g C_2 V_g^2 + Q^2}{2C_\Sigma}. \tag{6.26}$$

The free energy corresponds to the difference of the electrostatic energy and the work done in delivering charges from the source to the SET system, $F(n_1, n_2) = E_c - W$. The work done in delivering charge to the system is given by the time integral of the power delivered to the SET from the external sources as

$$W = \sum_{\text{all sources}} I(t)V(t)\mathrm{d}t = V_a \Delta Q_a + V_g \Delta Q_g, \tag{6.27}$$

where $\Delta Q_{a,g}$ is the total charge transferred from the drain or gate voltage sources, including the integer number of electrons that tunnel into and out of the island, as well as the continuous polarization charge that builds up in response to the change in electrostatic potential on the island. We can now look at the change in free energy of the entire circuit due to electrons tunneling across junctions 1 and 2 separately by considering the total free energy before and after tunnel events which decrease or increase the net number of electrons tunneling across junctions 1 and 2 as

$$\begin{aligned} \Delta F_1^{\pm} &= F(n_1 \pm 1, n_2) - F(n_1, n_2) \\ &= \frac{e}{C_\Sigma}\left(\frac{e}{2} \pm [(C_2 + C_g)V_a - C_g V_g + ne - Q_0]\right), \end{aligned} \tag{6.28}$$

$$\Delta F_2^{\pm} = F(n_1, n_2 \pm 1) - F(n_1, n_2) = \frac{e}{C_\Sigma}\left(\frac{e}{2} \pm [C_1 V_a + C_g V_g - ne + Q_0]\right). \tag{6.29}$$

We can now argue that the only high likelihood tunneling events are those that result in transitions to final states of lower energy, i.e., negative change in the free energy above.

6.5.2 Coulomb Blockade

Assume for simplicity that we have a double junction system without an external gate, i.e., $C_g = 0$ and that the stray polarization charge is zero.

Furthermore, assume that the two tunnel junctions are identical, i.e., $C_1 = C_2 = C$. If we start from a condition in which the island is initially charge neutral, i.e., $n = 0$, then it is clear from Eqs. (6.27) and (6.28) that there is a minimum applied drain voltage, V_a, necessary in either direction before the change in free energy is negative, i.e., for $-e/C_\Sigma < V_a e/C_\Sigma$, tunneling cannot occur. This phenomena is referred to as Coulomb blockade (CB), or the suppression of tunneling due to the effective charging energy barrier to adding or removing an electron from the island. An illustration of this phenomenon in terms of the energy band diagram of the system and the expected *I–V* characteristics are shown in Fig. 6.33.

Essentially, the Coulomb charging energy opens a gap in the continuous spectrum of energy states associated with the island, which forbids tunneling until this barrier is surmounted with an applied bias. Once an electron enters the dot (i.e., $n = 1$), a new Coulomb blockade exists until the electron tunnels out the other side. Hence, tunneling in this idealized situation also corresponds to the *correlated tunneling* of one electron at a time for biases just above this threshold. Higher thresholds exist in which it is energetically favorable for two electrons, three electrons, etc., to be injected, which for asymmetric barriers results in a *Coulomb staircase* corresponding to a series of

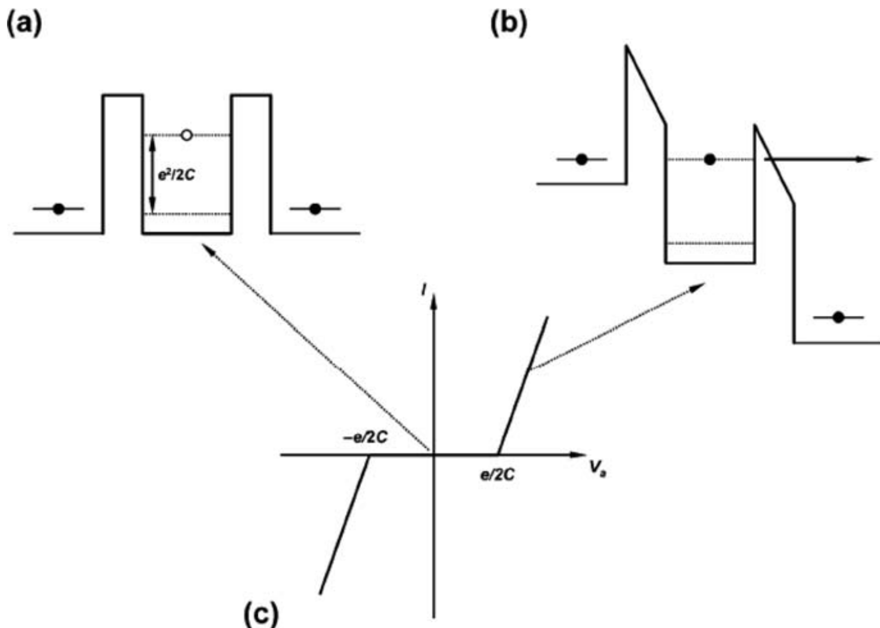


Fig. 6.33 Energy band diagram for a double junction systems illustrating the band diagram in the (**a**) Coulomb blockade regime and (**b**) tunneling regime. (**c**) Coulomb gap in the current–voltage characteristic for $-e = 2C < V < e = 2C$

current plateaus of successively high integer numbers of electrons tunneling across the double junction.

6.5.2.1 Coulomb Oscillations and the Single Electron Transistor

If we now consider the effect of the gate capacitance and bias in Eqs. (6.27) and (6.28), we see that the CB may be lifted with appropriate combination of positive or negative gate bias and number of excess electrons on the dot, n. Therefore, as a function of gate and source–drain bias, there are going to be regions where the free energy change is positive, corresponding to little current flow, and regions where tunneling is allowed energetically. This may be conveniently represented by a *stability diagram* as shown in Fig. 6.34. There the shaded regions correspond to combinations of the two biases where CB occurs, which forms the diamond pattern shown for integral values of the electron number on the dot. For successive changes of the effective gate charge, $C_g V_g = e$, the source–drain conductance goes through successive oscillations or resonances where CB is lifted.

This oscillatory behavior can be better understood looking at the energy band diagram on and off resonance as shown in Fig. 6.35, and the corresponding current–voltage characteristics. The effect of the gate is to tune the Coulomb gap in the density of states through the Fermi energies on the left and right (which are nearly coincident for small source–drain bias). As the gap is pulled below the Fermi energies, electron tunneling onto the island can occur, increasing the number n by one, and resulting in a new CB regime, hence the successive diamonds along the V_g axis. The corresponding conductance then exhibits a series of peaks spaced periodically in $\Delta V_g = e/C_g$, which are sometimes referred to as Coulomb oscillations.

Figure 6.35 is relevant for the idealized case of a perfectly conducting metallic island consisting of many electrons with quasi-continuous energy spectrum. In the case that the island is a semiconductor dot formed by artificial

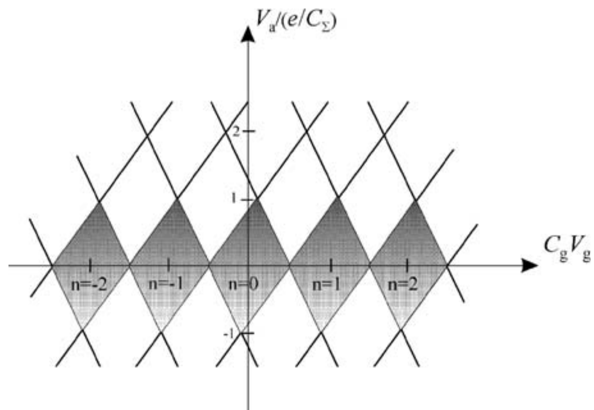


Fig. 6.34 Stability diagram for the single electron transistor of Fig. 6.33

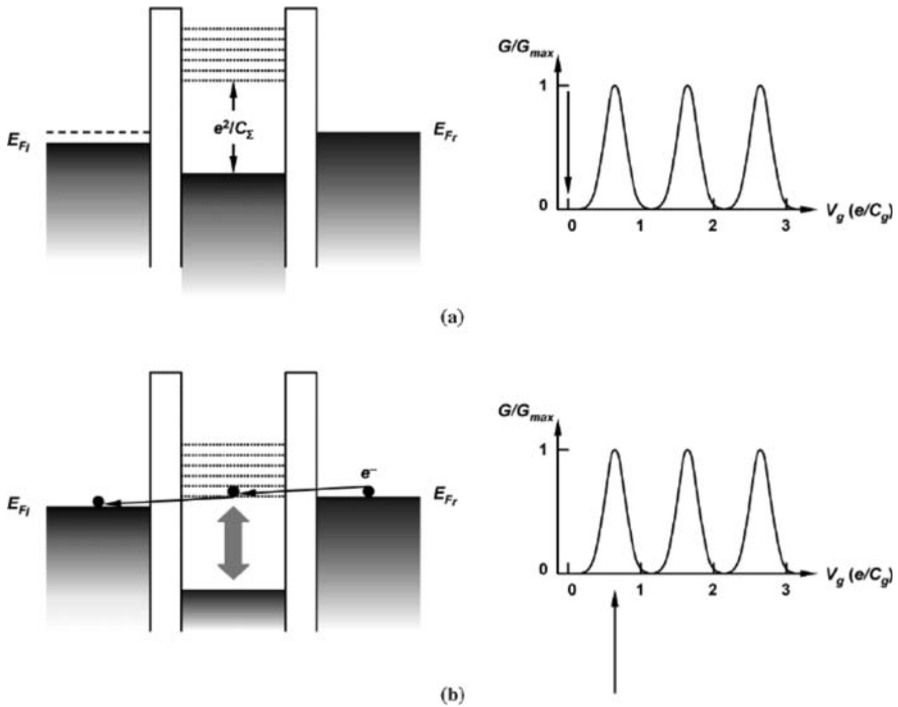


Fig. 6.35 Energy band diagram and *I–V* characteristics as a function of gate bias of a SET transistor under small source–drain bias (**a**) in the Coulomb blockade regime and (**b**) in resonance

confinement, the quantized energy of the dot states can be appreciable compared to the Coulomb energy and must be accounted for in the description. If we label the energy levels in the dot as E_n, then successive oscillations in the conductance with gate bias include both the Coulomb contribution and the energy spacing of successive quantized levels:

$$\Delta V_{\mathrm{g}} = \frac{C_\Sigma}{C_{\mathrm{g}}}\left(\frac{E_{n+1} - E_n}{e}\right) + \frac{e}{C_{\mathrm{g}}}. \tag{6.30}$$

Since in general the energy spacing of the quantized levels of an artificial molecule are non-uniform, the overall spacing of the Coulomb oscillations with gate bias will no longer be strictly uniform.

Within the picture presented above of Coulomb blockade and Coulomb oscillations, we have implicitly assumed that the electrons are well localized on the dot, i.e., that we can talk about the electron as residing inside or outside the dot. Hence, the probability of tunneling in and out of the dot should be sufficiently large that the electron is localized on the dot, which is

usually satisfied if the tunneling resistance itself is much larger than the fundamental resistance corresponding to a single conducting channel, i.e., $R >> h/e^2$.

6.5.3 SET Modeling and Simulation

To actually formulate the current–voltage characteristics of SET structures, a kinetic equation approach was generalized by Averin and Likharev [85] now referred to as the 'orthodox' theory of single electron tunneling. Within a kinetic equation approach, tunneling processes are considered as random scattering events that instantaneously change the energy of the system. Using time-dependent perturbation theory, the usual theory of tunneling via the tunneling Hamiltonian approach (which treats tunneling via perturbation theory) can be generalized to include the change in free energy discussed above before and after tunneling. Hence the tunneling rate for the *j*th tunnel junction in an N junction system is given by

$$\Gamma_j^{\pm}(n) = \frac{1}{R_j e}\left(\frac{\Delta F_j^{\pm}/e}{1 - \exp\left(-\Delta F_j^{\pm}/k_B T\right)}\right), \tag{6.31}$$

where ΔF_j is the change in free energy, which as defined for the two junction SET system by Eqs. (6.27) and (6.28). Within the kinetic equation framework, we can define the distribution function for the *island* occupancy, $f(n_1, n_2, n_3, \ldots, n_{N-1})$, which is the probability of the system having n_1 electrons on island 1, n_2 electrons on island 2, etc. A kinetic or master equation can then be derived which represents an equation of motion for f through a detailed balance of tunneling events onto and off each island:

$$\begin{aligned}\frac{\partial f(n_1, n_2, \ldots, t)}{\partial t} = \sum_{j=1,N} \Big\{ & \Gamma_j^{+}(n_1, \ldots, n_j - 1) f(n_1, \ldots, n_j - 1 \ldots t) \\ & + \Gamma_j^{-}(n_1, \ldots, n_j + 1) f(n_1, \ldots, n_j + 1 \ldots t) \\ & - \left[\Gamma_j^{+}(n_1, \ldots, n_j) + \Gamma_j^{-}(n_1, \ldots, n_j)\right] f(n_1, \ldots, n_j \ldots t) \Big\},\end{aligned} \tag{6.32}$$

where here $\Gamma_j^{\pm}$ refers to the net tunneling onto each island from all possible junctions using the junction rate defined in Eq. (6.30). Once f is calculated, averages may be calculated for quantities of interest such as the total energy, or current flow through a particular junction. Direct solution of this master equation has been successfully used to model the *I–V* characteristics of the single electron transistor discussed in the previous section (see, for example, [18]).

The derivation of Eq. (6.31) is based on first-order time-dependent perturbation theory (i.e., Fermi's golden rule); however, higher order tunneling processes may in fact be important, particularly when the tunnel resistance approaches that of the fundamental conductance, h/e^2.Higher order processes, or *co-tunneling*, represent tunneling processes that occur through multiple junctions, such as, for example, resonant tunneling in a double barrier resonant tunneling diode. The theory of co-tunneling has been developed by Averin and Nazarov [86], which gives corrections to second order in the inverse tunnel resistance, and gives rise to additional power law dependencies on voltage and temperature. The tunnel resistance dependence has been studied in detail experimentally using quantum point contact structures (where the tunnel resistance can be tuned) [87].

While direct solution of the master equation (6.31) is feasible, for arbitrary large SET circuits, it has become increasingly popular to utilize Monte Carlo techniques for the simulation of single electron tunneling [84, 88, 89, 23]. In Monte Carlo simulation, basically the stochastic tunneling events across all possible junctions based on Eq. (6.30) (and extensions to higher order and nonlinear tunneling resistances) are simulated in time using the computer random number generator to generate the time between tunneling events. Commercial simulators are available such as SIMON [90] which provide schematic capture for design and simulation of single electron circuits.

6.5.4 Recent Experimental Studies

An exhaustive review of experimental work on single electron phenomena is beyond the scope of this chapter. Basically, Coulomb blockade and associated single electron behavior such as the Coulomb staircase were first observed in metal–oxide tunnel junction systems in the 1980s (see [83] and references therein). As mentioned earlier, Fulton and Dolan [85] fabricated the first successful single electron transistor, in which the CB regime and Coulomb staircase could be controllably modified by a gate. Following this, researchers were able to realize single electron turnstiles and pumps [91, 21] in which single electrons could be systematically clocked through an array of tunnel junctions by periodic modulation of the gate potential at rf frequencies, and the resulting current is given quite accurately by $I = ef$, where f is the ac frequency. Such turnstile devices are still an active area of investigation for accurate metrological standards.

The first definitive demonstration of Coulomb blockade in semiconductor structures was reported by Meirev et al. using a pair quantum point contact structures to form a double tunnel junction system over a high-mobility 2 DEG GaAs/AlGaAs heterostructure to form a quantum dot as the island, and using substrate bias as the gate potential [92]. Clear periodic oscillations of the source–drain conductance with gate bias were observed. Subsequent

demonstration of the Coulomb staircase and turnstile behavior in more elaborate gate geometry QPC quantum dots in high-mobility 2 DEG material were reported by the Delft group [93]. The flexibility of the QPC structure has led to increasingly more complicated geometries to investigate single electron tunneling through multiple dots, where molecular 'hybridization' of the states in coupled dots is observed when the dots are allowed to interact (see, for example, [94]).

6.5.4.1 Si Nanoelectronic Devices

More recently, interest has been focused on the development of single electron devices in Si that are compatible with Si CMOS processing for potential realization of Si nanoelectronic circuits. The advantages of such Si-based nanoelectronic systems is that Si process technology is more mature, and the high quality of the native oxide as well as improvements in Si on insulator (SOI) technology provides more flexibility in design of single electron devices than III–V compound technologies. The main disadvantage is the higher effective mass, lower mobility, and generally higher density of defects in Si compared to high-purity epitaxial growth techniques used for the 2 DEG structures discussed earlier, which tend to mask quantum and single electron behavior.

High-resolution nanofabrication techniques such as STM/AFM lithography have led to considerable reduction in single electron devices. Sub-10 nm tunnel junctions have been realized using in situ anodization of Ti films with an AFM tip [95]. Similar technology was used by Matsumoto to fabricate ultrasmall double barrier SET structures of anodized Ti on an oxidized Si substrate, which exhibited strong evidence of Coulomb staircase with 150 mV period at room temperature [96].

Conventional CMOS technology has been used to realize single electron structures as well [97, 98, 99]. Figure 6.36 shows the schematic of a split-gate quantum point contact structure similar to those discussed already, but fabricated in a double oxide, Si MOS transistor structure [98]. As shown in Fig. 6.36b, distinct Coulomb oscillations are observed at 4 K as a function of the top inversion gate bias, superimposed on a background of rising conductance as the channel forming the QPCs opens up with increasing gate bias. A third terminal or plunger adjacent to the dot acts like the gate of a SET, while at the same time changing the shape of the dot itself. A plot of the location of the peak conductance as a function of both plunger is shown in Fig. 6.37 as well. As can be seen, there are several sets of peaks that evolve with different slope, and which exhibit anti-crossing behavior which appears somewhat analogous to that of atomic levels. In fact, it appears that in these dots, conductance oscillations are dominated by the energy spectrum of the dot itself as much as by the Coulomb charging energy as given by Eq. (6.29).

Silicon on insulator technology (SOI) has gained increasing acceptance in recent years as a technology for scaling conventional CMOS technology below the 10 nm gate length node. SOI technology has also proved promising in

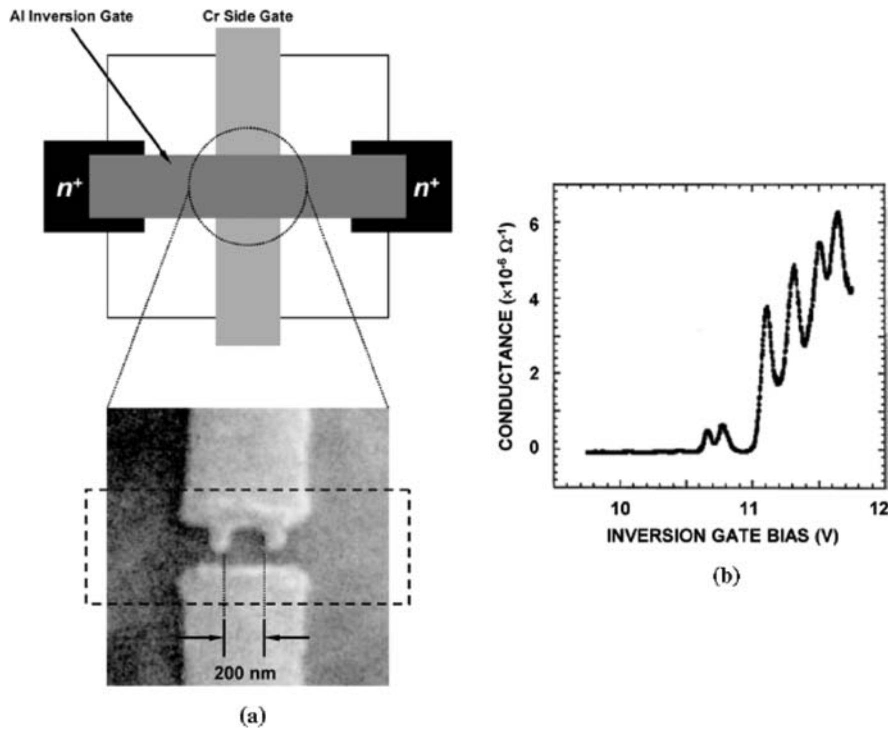


Fig. 6.36 MOS single electron transistor. (**a**) Illustration of the double oxide split-gate structure and (**b**) associated conductance–voltage characteristics with inversion layer bias [98]

realizing SET structures with potential for room temperature operation. Zhuang et al. first demonstrated a quantum dot structure fabricated on an SOI wafer [100], in which the SOI layer is etched down forming a corrugated Si wire over which a poly-Si gate is deposited.

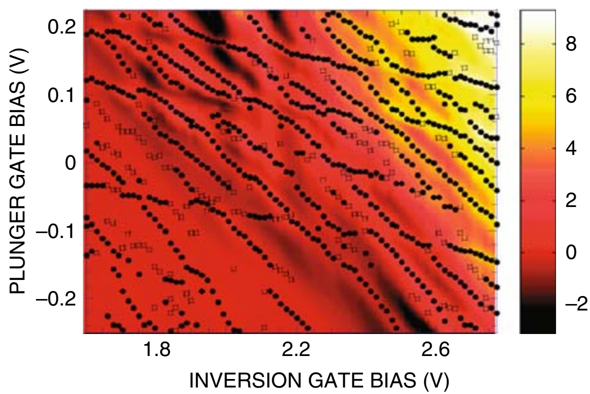


Fig. 6.37 Conductance peak positions as function of both inversion gate and plunger gate bias exhibiting crossing and anti-crossing behaviors of apparent level structure of dot [103] (*See* Color Insert)

6.5.4.2 Coulomb Blockade in Self-Assembled Systems

Self-assembled structures such as CNTs and NWs are natural candidates for observing single electron phenomena due to their inherently small diameters, and correspondingly small capacitances. For the case of CNTs, the high conductivity of the CNT itself serves to form a Coulomb island, with tunnel barriers into and out of the island formed from Schottky contacts to either end of the nanotube. Tan and co-workers [101] observed clear evidence of Coulomb oscillations and Coulomb staircase behavior in transport studies of single-wall CNTs.

Single electron transistors may be fabricated in semiconductor nanowires by growing thin tunnel barriers epitaxially [50]. Figure 6.38 shows the growth and electron micrograph of a double barrier structure grown using chemical beam epitaxy (CBE), where thin InP barriers of various spacings are grown in a smaller bandgap InAs wire [102]. The wires are

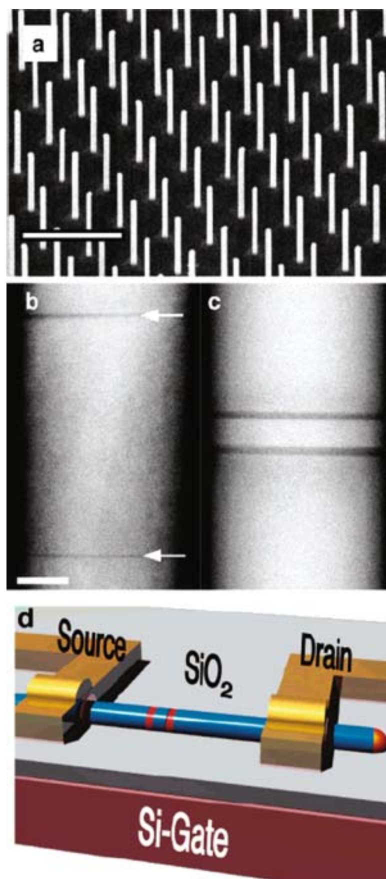


Fig. 6.38 Characterization and processing of nanowires. (**a**) Scanning electron micrograph of homogeneous InAs nanowires grown on an InAs substrate from lithographically defined arrays of Au particles. The image demonstrates the ability of the CBE to produce identical nanowire devices. The scale bar corresponds to 1 μm. (**b**) Dark-field scanning transmission electron microscopy image of a nanowire with a 100 nm long InAs quantum dot between two very thin InP barriers. Scale bar depicts 20 nm. (**c**) Corresponding image of a 10 nm long InAs dot. The InP barrier thickness is 3 and 3.7 nm, respectively. (**d**) The heterostructured wires are deposited on a SiO_2-capped Si substrate and source and drain contacts are fabricated by lithography. Reprinted with permission from Ref. [102], Copyright 2004, American Chemical Society (*See* Color Insert)

dispersed on a Si substrate with an oxide to form a backgate as shown schematically in Fig. 6.38c.

Figure 6.39 illustrates nicely the dependence of the observed oscillations on the size of the Coulomb island defined by the distance between the two barriers. For large distances (Fig. 6.39a), the oscillations are periodic, and the device behaves like a metallic SET. For spacings less than 30 nm, the Coulomb charging energy is comparable to the level separation due to

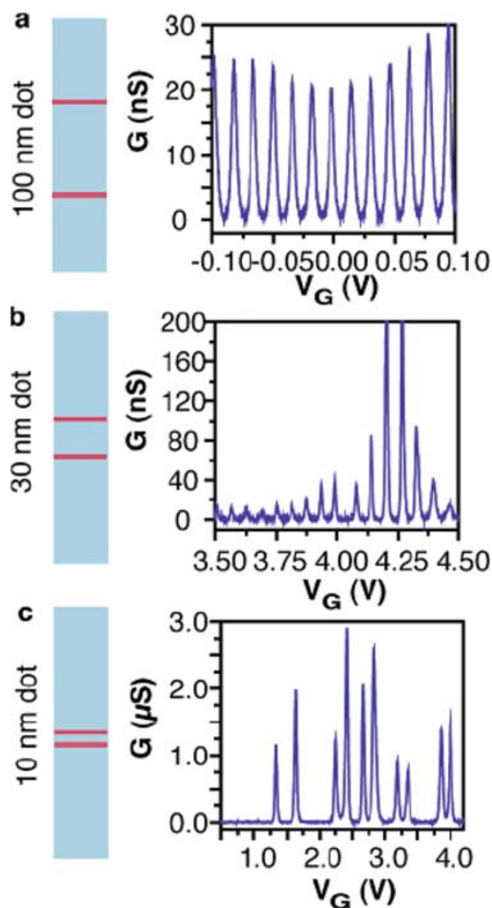


Fig. 6.39 Effect of reduced quantum dot length. (**a**) Gate characteristics of a SET with a 100 nm long dot. The oscillations are perfectly periodic and are visible up to 12 K. (**b**) When the dot length is 30 nm, the level spacing at the Fermi energy is comparable to the charging energy and the Coulomb oscillations are no longer completely periodic. (**c**) A 10 nm dot results in a device depleted of electrons at zero gate voltage. By increasing the electrostatic potential electrons are added one by one. For some electron configurations, the addition energy is larger corresponding to filled electron shells. All data in this figure were recorded at 4.2 K. Reprinted with permission from Ref. [102], Copyright 2004, American Chemical Society

confinement by the barriers, and period of oscillations is a combination of level spacing and Coulomb charging energy as described by Eq. (6.29). At the narrowest spacing, the energy spacing is dominated by the molecular states of the dot, leading to a non-periodic structure.

6.6 Summary

In this review, we have discussed some of the basic transport phenomena occurring in structures at the nanoscale. As mentioned in the introduction, the current drive toward nanoelectronic technologies is driven both by top-down scaling of dimensions in semiconductor transistors and by bottom-up self-assembly of structures such as carbon nanotubes, semiconductor, and metallic nanowires and nanocrystals. We discussed the transition from semi-classical diffusive transport at mesoscopic to macroscopic scales, to fully coherent quantum transport as characteristic dimensions are reduced below the mean free paths for scattering in the system. Effects such as quantum confinement and reduced dimensionality of carriers, quantum mechanical transmission and reflection, and single electron effects such as Coulomb blockade, all become manifest as dimensions reduce, giving rise to new paradigms of charge transport, and interesting new ideas for functional devices and architectures which may provide alternatives for current technology as fundamental limits are reached.

References

1. G. Binnig and H. Rohrer, *Appl. Phys. Lett.* **40**, 178 (1982).
2. 2006 International Technology Roadmap of Semiconductors, http://public.itrs.net/
3. R. Chau, S. Datta, M. Doczy, B. Doyle, B. Jin, J. Kavalieros, A. Majumdar, M. Metz, and M. Radosavljevic, *IEEE Trans. Nanotechnol.* **4**, 153 (2005).
4. D. K. Ferry and S. M. Goodnick, *Transport in Nanostructures*, Cambridge University Press, Cambridge, 1997.
5. M. Depas, B. Vermeire, P. W. Mertens, R. L. Van Meirhaegne, and M. M. Heyns, *Solid-State Electron.* **38**, 1465 (1995).
6. M. Leong, H.-S. Wong, E. Nowak, J. Kedzierski, and E. Jones, *Proceedings of the International Symposium on QED2002*, 492 (2002).
7. B. J. van Wees, H. van Houten, C. W. J. Beenakker, J. G. Williamson, L. P. Kouwenhoven, D. van der Marel, and C. T. Foxon, *Phys. Rev. Lett.* **60**, 848 (1988).
8. D. A. Wharam, T. J. Thornton, R. Newbury, M. Pepper, H. Ahmed, J. E. F. Frost, D. G. Hasko, and D. C. Peacock, *J. Phys. C* **21**, L209 (1988).
9. S. Washburn, in *Mesoscopic Phenomena in Solids*, B. L. Altshuler, P. A. Lee, and R. A. Webb (eds.) (Elsevier, North-Holland, Amsterdam, 1991) pp. 1–36.
10. R. E. Prange and S. M. Girvin (eds.) *The Quantum Hall Effect*, 2 nd Edition (Springer-Verlag, New York, 1990)
11. F. Sols, M. Macucci, U. Ravaioli, and K. Hess, *J. Appl. Phys.* **66**, 3892 (1989).
12. S. Datta, *Superlatt. Microstruct.* **6**, 83 (1989).

13. A. Weisshaar, J. Lary, S. M. Goodnick, and V. K. Tripathi, *Appl. Phys. Lett.* **55**, 2114 (1989).
14. L. Worschech, B. Weidner, S. Reitzenstein, and A. Forchel, *Appl. Phys. Lett.* **78**, 3325 (2001).
15. K. Hieke and M. Ulfward, *Phys. Rev. B* **62**, 16727 (2000).
16. K. K. Likharev, *Proc. IEEE* **87**, 606 (1999).
17. *Single Charge Tunneling, Coulomb Blockade Phenomena in Nanostructures*, H. Grabert and M. H. Devoret (eds.) NATO ASI Series B 294 (Plenum Press, New York, 1992).
18. K. Likharev, *IBM J. Res. Dev.* **32**, 144 (1988).
19. L. J. Geerligs, V. F. Anderegg, P. A. M. Holweg, J. E. Mooij, H. Pothier, D. Esteve, C. Urbina, and M. H. Devoret, *Phys. Rev. Lett.* **54**, 2691 (1990).
20. L. P. Kouwenhouven, A. T. Johnson, N. C. van der Vaart, C. J. P. M. Harmans, and C. T. Foxon, *Phys. Rev. Lett.* **67**, 1626 (1991).
21. H. Pothier, P. Lafarge, C. Urbina, D. Esteve, and M. H. Devoret, *Europhys. Lett.* **17**, 249 (1992).
22. D. H. Kim, S.-K. Sung, K. R. Kim, J. D. Lee, B.-G. Park, B. Ho Choi, S. W. Hwang, and D. Ahn, *IEEE Trans. ED* **49**, 627 (2002).
23. C. Wasshuber, H. Kosina, S. Selberherr, *IEEE Trans. CAD* **16**, 937 (1997).
24. Y. Cui and C. M. Lieber, *Science* **291**, 851 (2001).
25. R. Martel, V. Derycke, C. Lavoie, J. Appenzeller, K. K. Chan, J. Tersoff, and Ph. Avouris, *Phys. Rev. Lett.* **87**, 256805 (2001).
26. M. A. Reed, J. N. Randall, R. J. Aggarwal, R. J. Matyi, T. M. Moore, A. E. Wetsel, *Phys. Rev. Lett.* **60**, 535 (1988).
27. L. Zhuang, L. Guo, and S. Y. Chou, *Appl. Phys. Lett.* **72**, 1205 (1998).
28. D. H. Kim, S.-K. Sung, K. R. Kim, J. D. Lee, B.-G. Park, B. Ho Choi, S. W. Hwang, and D. Ahn, *IEEE Trans. ED* **49**, 627 (2002).
29. K. Hiruma, M. Yazawa, T. Katsuyama, K. Haraguchi, M. Koguchi, and H. Kakibayashi, *J. Appl. Phys.* **77**, 447 (1995).
30. H. Dai, E. W. Wong, Y. Z. Lu, S. Fan, and C. M. Lieber, *Nature* **375**, 769 (1995).
31. Y. Cui, X. Duan, J. Hu, and C. M. Lieber, *J. Phys. Chem. B* **104**, 5213 (2000).
32. M. T. Björk, B. J. Ohlsoon, T. Sass, A. I. Persson, C. Thelander, M. H. Magnusson, K. Deppert, L. R. Wallenbeg, and L. Samuelson, *Appl. Phys. Lett.* **80**, 1058 (2002).
33. M. T. Björk, B. J. Ohlsoon, T. Sass, A. I. Persson, C. Thelander, M. H. Magnusson, K. Deppert, L. R. Wallenbeg, and L. Samuelson, *Nano Lett.* **2**, 87 (2002).
34. M. S. Dresselhaus, G. Dresselhaus, and P. C. Eklund, *Science of Fullerenes and Carbon Nanotubes*, Academic Press, Inc., New York (1996).
35. D. K. Ferry, *Semiconductors*, Macmillan, New York, 1991.
36. D. Vasileska and S. M. Goodnick, “Computational Electronics,” *Mater. Sci. Eng. Rep.* **R38**, 181 (2002).
37. L. I. Schiff, *Quantum Mechanics*, McGraw-Hill Inc., New York, 1955.
38. C. Jacoboni and P. Lugli, *The Monte Carlo Method for Semiconductor Device Simulation*, Springer-Verlag, Vienna, 1989.
39. S. Yamakawa, S. Aboud, M. Saraniti, and S. M. Goodnick, *Semicond. Sci. Technol.* **19**, S475 (2004).
40. T. Ando, A. B. Fowler, and F. Stern, *Rev. Mod. Phys.* **54**, 437 (1982).
41. G. Bastard, J. A. Brum, and R. Ferreira, *Solid State Phys.* **44**, 437 (1982)
42. R. Dingle, H. Störmer, A. C. Gossard, and W. Wiegmann, *Appl. Phys. Lett.* **33**, 665 (1978).
43. L. Pfeiffer et al., *Appl. Phys. Lett.* **55**, 1888 (1989).
44. B. J. F. Lin, D. C. Tsui, M. A. Paalanen, and A. C. Gossard, *Appl. Phys. Lett.* **45**, 695 (1984).
45. K. Hiruma, M. Yazawa, T. Katsuyama, K. Haraguchi, M. Koguchi, and H. Kakibayashi, *J. Appl. Phys.* **77**, 447 (1995).

46. H. Dai, E. W. Wong, Y. Z. Lu, S. Fan, and C. M. Lieber, *Nature* **375**, 769 (1995).
47. Y. Cui, Z. Zhong, D. Wang, W. U. Wang, and C. M. Lieber, *Nano Lett.* **3**, 149 (2003).
48. X. Duan, C. Niu, V. Sahl, J. Chen, J. W. Parce, S. Empedocies, and J. L. Goldman, *Nature* **425**, 274 (2003).
49. M. T. Björk, B. J. Ohlsoon, C. Thelander, A. I. Persson, K. Deppert, L. R. Wallenbeg, and L. Samuelson, *Appl. Phys. Lett.* **81**, 4458 (2002).
50. C. Thelander, T. Martensson, M. T. Björk, B. J. Ohlsson, M. W. Larsson, L. R. Wallenberg, and L. Samuelson, *Appl. Phys. Lett.* **83**, 2052 (2003).
51. Z. Zhong, D. Wang, Y. Cui, M. W. Bockrath, and C. M. Lieber, *Science* **302**, 1377–1379 (2003).
52. P. L. McEuen, M. S. Fuhrer, and H. Park, *IEEE Trans. Nanotechnol.* **1**, 78 (2002)
53. E. B. Ramayya, D. Vasileska, S. M. Goodnick, and I. Knezevic, *IEEE Trans. Nanotechnol.* **6**, 113 (2007).
54. H. Majima, H. Ishikuro, and T. Hiramoto, *IEEE Electron Device Lett.* **21**, 396 (2000).
55. E. B. Ramayya, D. Vasileska, S. M. Goodnick, and I. Knezevic, *J. Comput. Electron.* accepted for publication (2008).
56. S. Iijama, *Nature* **363**, 603 (1993).
57. T. Dürkop, S. A. Getty, E. Cobas, and M. S. Fuhrer, *Nano Lett.* **4**, 35 (2004).
58. A. Javey, J. Guo, Q. Wang, M. Lundstrom, and H. Dai, *Nature* **424**, 654 (2003).
59. L. Balents and M. P. A. Fisher, *Phys. Rev. B* **55**, 11973 (1997).
60. C. L. Kane and E. J. Mele, *Phys. Rev. Lett.* **78**, 1932 (1997).
61. Z. Yao, C. L. Kane, and C. Dekker, *Phys. Rev. Lett.* **84**, 2941 (2000).
62. R. Tsu and L. Esaki, *Appl. Phys. Lett.* **22**, 562 (1973).
63. L. L. Chang, L. Esaki, and R. Tsu, *Appl. Phys. Lett.* **24**, 593 (1974).
64. T. C. L. G. Sollner, W. D. Goodhue, P. E. Tannenwald, C. D. Parker, and D. D. Peck, *Appl. Phys. Lett.* **43**, 588 (1983).
65. S. Luryi, *Appl. Phys. Lett.* **47**, 490 (1985).
66. R. Landauer, *IBM J. Res. Dev.* **1**, 223 (1957).
67. R. Landauer, *Philos. Mag.* **21**, 863 (1970).
68. R. Büttiker, Y. Imry, R. Landauer, and S. Pinhas, *Phys. Rev. B* **31**, 6207 (1985).
69. K. J. Thomas, J. T. Nicholls, N. J. Appleyard, M. Y. Simmons, M. Pepper, D. R. Mace, W. R. Tribe, and D. A. Ritchie, *Phys. Rev. B* **58**, 4846 (1998).
70. G. Timp, *Semiconductors and Semimetals*, vol. 35, pp. 113–190, M. A. Reed (ed.) (Academic Press, New York, 1992).
71. J. C. Wu, M. N. Wybourne, A. Weisshaar, and S. M. Goodnick, *J. Appl. Phys.* **74**, 4590 (1993).
72. T. Palm and L. Thylén, *Appl. Phys. Lett.* **60**, 237 (1992).
73. T. Palm, *Phys. Rev. B* **52**, 13773 (1995).
74. J.-O. J Wesström, *Phys. Rev. Lett.* **82**, 2564 (1999).
75. H. Q. Xu, *Appl. Phys. Lett.* **78**, 2064 (2001).
76. H. Q. Xu, *Appl. Phys. Lett.* **80**, 853 (2002).
77. K. Hieke and M. Ulfward, *Phys. Rev. B* **62**, 16727 (2000).
78. L. Worschech, H. Q. Xu, A. Forchel, and L. Samuelson, *Appl. Phys. Lett.* **79**, 3287 (2002).
79. I. Shorubalko, H. Q. Xu, I. Maximov, P. Omling, L. Samuelson, and W. Seifert, *Appl. Phys. Lett.* **79**, 1384 (2001).
80. S. Kasai and H. Hasegawa, *IEEE Electron Device Lett.* **23**, 446 (2002).
81. H. Grabert and M. H. Devoret (eds.) *Single Charge Tunneling, Coulomb Blockade Phenomena in Nanostructures*, NATO ASI Series B 294 (Plenum Press, New York, 1992).
82. C. Wasshuber, *Computational Single-Electronics*, Springer, New York, 2001.
83. T. A. Fulton and G. J. Dolan, *Phy. Rev. Lett.* **59**, 109 (1987).
84. L. S. Kuz'min and K. K. Likharev, *JETP Lett.* **45**, 495 (1987).
85. D. V. Averin and K. K. Likharev, *Single Electronics: A Correlated Transfer of Single Electrons and Cooper Pairs in Systems of Small Tunnel Junctions*, B. L. Altshuler, P. A. Lee,

and R. A. Webb (eds.) *Mesoscopic Phenomena in Solids*, pp. 173–271 (Amsterdam, Oxford, New York, Tokyo, 1991).
86. D. V. Averin and Yu. V. Nazarov, *Phys. Rev. Lett.* **65**, 2446 (1990).
87. C. Pasquier, U. Meirav, F. I. B. Williams, D. G. Glattli, Y. Jin, and B. Etienne, *Phys. Rev. Lett.* **70**, 69 (1993).
88. R. H. Chen, A. N. Korotkov, and K. K. Likharev, *Appl. Phys. Lett.* **68**, 1954 (1996).
89. M. Kirihara, N. Kuwamura, K. Taniguchi, and C. Hamaguchi, *Proceedings of the International Conference on Solid State Devices and Materials*, Yokohama, 1994, pp. 328–330.
90. C. Wasshuber and H. Kosina, *Superlattices and Microstructures* **21**, 37 (1997).
91. L. J. Geerligs, V. F. Anderegg, P. A. M. Holweg, J. E. Mooij, H. Pothier, D. Esteve, C. Urbina, and M. H. Devoret, *Phys. Rev. Lett.* **54**, 2691 (1990)
92. U. Meirav, M. A. Kastner, and S. J. Wind, *Phys. Rev. Lett.* **65**, 771 (1990); M. A. Kastner, *Rev. Mod. Phys.* **64**, 849 (1992).
93. L. P. Kouwenhouven, A. T. Johnson, N. C. van der Vaart, C. J. P. M. Harmans, and C. T. Foxon, *Phys. Rev. Lett.* **67**, 1626 (1991); *Zeitschrift Physik B* **85**, 381 (1991).
94. F. R. Waugh, M. J. Berry, D. J. Mar, R. M. Westervelt, K. L. Campman, and A. C. Gossard, *Phys. Rev. Lett.* **75**, 705 (1995).
95. E. S. Snow and P. M. Cambell, *Science* **270**, 1639 (1995).
96. K. Matsumoto, *Phys. B* **227**, 92 (1996).
97. H. Matsuoka and S. Kimura, *Appl. Phys. Lett.* **66**, 613 (1995).
98. M. Khoury, M. J. Rack, A. Gunther, and D. K. Ferry, *Appl. Phys. Lett.* **74**, 1576 (1999); A. Gunther, M. Khoury, S. Milicic, D. Vasileska, T. Thornton, and S. M. Goodnick, *Superlatt. Microstruct.* **27**, 373 (2000).
99. F. Simmel, D. Abusch-Magder, D. A. Wharam, M. A. Kastner, and J. P. Kotthaus, *Phys. Rev. B* **59**, R10441 (1999).
100. L. Zhuang, L. Guo, and S. Y. Chou, *Appl. Phys. Lett.* **72**, 1205 (1998).
101. S. J. Tans, M. H. Devoret, H. Dai, A. Thess, R. E. Smalley, L. J. Geerligs, and C. Dekker, *Nature* **386**, 474 (1997).
102. M. T. Björk, C. Thelander, A. E. Hansen, L. E. Jensen, M. W. Larsson, L. R. Wallenberg, and L. Samuelson, *Nano Lett.* **4**, 1621–1625 (2004).

Chapter 7
Density Functional Theory of High-*k* Dielectric Gate Stacks

Alexander A. Demkov

Abstract Density functional theory has proved to be a useful tool in device engineering, particularly at nanoscale and when novel materials are involved. In this chapter we briefly introduce the theoretical background necessary for understanding the modern theory of solid state and review recent theoretical results in the area of advanced gate stack materials engineering.

7.1 Introduction

As scaling of the complementary metal oxide semiconductor (CMOS) technology takes us below 65 nm many new materials, traditionally not associated with the semiconductor process, are being introduced into manufacturing. Notably, transition metal (TM) oxides or more generally dielectrics with a high dielectric constant or high-*k* dielectrics are being considered for the gate stack applications instead of SiO_2. The gate stack is a multilayer structure in place of the metal oxide semiconductor capacitor (see Fig. 7.1). Its capacitance controls the saturation current and has been traditionally maintained by reducing its thickness in accord with the gate length reduction (the so-called scaling at the heart of Moor's law). However, after reaching the oxide thickness of 12 Å the scaling has more or less stopped due to the prohibitively large gate leakage current caused by direct tunneling across the gate oxide. Thus a new dielectric with a larger dielectric constant has to be introduced. After the introduction of Cu this is arguably the most drastic departure from the traditional CMOS process. The physics and chemistry of these materials is much more complicated than that of Si_3N_4 or SiO_2, and theoretical calculations of their properties have proven to be extremely useful in both process development and device engineering. The work horse of the modern computational materials science is density functional theory (DFT) within the local density approximation (LDA) and pseudopotential (PP) approximation. In this

A.A. Demkov
Department of Physics, The University of Texas at Austin, Austin, TX 78712, USA
e-mail: demkov@physics.utexas.edu

A. Korkin, F. Rosei (eds.), *Nanoelectronics and Photonics*,
DOI: 10.1007/978-0-387-76499-3_7, © Springer Science+Business Media, LLC 2008

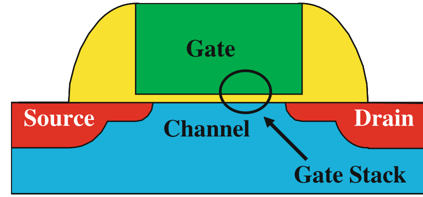


Fig. 7.1 Schematic of a field-effect transistor. The metal oxide semiconductor (MOS) capacitor is a multilayer structure known as the gate stack

chapter we shall review the basic concepts of this theoretical approach and give a brief overview of the recent results in the area of the theory of high-*k* dielectrics. The rest of the chapter is organized as follows. In Section 7.2 the DFT-LDA-PP scheme is briefly outlined. In Section 7.3 we discuss recent theoretical work on dielectric properties, defects, interfaces, band alignment and transport characteristics of gate stacks containing TM oxides. In Section 7.4 we briefly summarize our own results on the band alignment at Si/SiO_2, Si/HfO_2, SiO_2/HfO_2 and HfO_2/Mo interfaces.

7.2 Theoretical Background

7.2.1 Electrons and Phonons

Before discussing the density functional formalism used in most modern solid-state calculations I would like to outline the global landscape of the problem to put it into perspective. This discussion is intended for graduate students and can be omitted by the experts. The problem of describing solid state theoretically is its enormous complexity. A solid is comprised of electrons and nuclei interacting via Coulomb forces, so one has to describe correlated behavior of about 10^{23} particles! Clearly, this is an impossible problem unless some simplifications are made. The first step is to separate light and fast electrons from slow and heavy nuclei. The original idea belongs to Max Born and Robert Oppenheimer (Max Born was born in Breslau, Germany, in 1882; Robert Oppenheimer was born in New York in 1904) and was published in 1927 [1]. Note that Oppenheimer was only 23 years old when the paper came out. They suggested first to solve the electronic problem for some fixed configuration of nuclei $\bar{R}$:

$$\hat{H}_{\rm el}\varphi_i(\vec{r}_1,\vec{r}_2,\ldots;\bar{R}) = E_i^{\rm el}(\bar{R})\varphi_i(\vec{r}_1,\vec{r}_2,\ldots;\bar{R}) \tag{7.1}$$

It is customary to include the nucleus–nucleus (proton–proton) repulsion into the electronic Hamiltonian, so $\hat{H}_{\rm el}$ is given by

$$\hat{H}_{\rm el} = \hat{T}_{\rm e} + \hat{U}_{\rm ee} + \hat{U}_{\rm ep} + \hat{U}_{\rm pp} \tag{7.2}$$

Once we solve this problem we have a complete set of functions to expand the total (electrons and nuclei) wave function of the system:

$$\Psi_{\mathrm{s}}(\vec{r}_1, \vec{r}_2, \ldots, \vec{R}_1, \vec{R}_2 \ldots) = \sum_i \chi_i(\bar{R})\varphi_i(\vec{r}_1, \vec{r}_2, \ldots; \bar{R}) \tag{7.3}$$

Of course, the complete set we are using is changing all the time as the nuclei move, and in each particular case one needs to specify which configuration is used. This wave function is the Born–Oppenheimer ansatz. You insert this expression into Schrödinger equation for the entire system and average out fast electronic coordinates. This is achieved by multiplying the whole expression by the conjugate of the electronic wave function and integrating over all electronic coordinates. If we now neglect all terms involving derivatives with respect to nuclear positions with the exception of the nuclear kinetic energy, we end up having an effective Schrödinger-like equation for the coefficients χ which play the role of the nuclear wave functions:

$$\left[\hat{T}_{\mathrm{p}} + E_j(\bar{R})\right]\chi_{j\mathrm{s}}(\bar{R}) = E_{\mathrm{s}}\chi_{j\mathrm{s}}(\bar{R}) \tag{7.4}$$

The significance of this expression is that the potential energy of the nuclear motion is nothing more than the total electronic energy. We should not get carried way, however, for we still do not know how to solve the many-electron Schrödinger equation. In principle, the problem can be solved directly using the so-called quantum Monte Carlo methods, but in practice approximations are needed. Hartree–Fock theory is the simplest many-electron theory which essentially treats electrons as independent (the dynamic electron–electron interaction is handled in electrostatic approximation), but takes into account the Pauli principle. Unfortunately, this approximation does not describe solids very well. Density functional theory which we will now describe appears to do a better job.

7.2.2 Many-Electron Problem and Density Functional Theory

The modern electronic structure theory of materials is based on density functional theory introduced by Walter Kohn and co-workers in mid-1960s [2, 3]. The theory formulates the many-body problem of interacting electrons and ions in terms of a single variable, namely the electron density. The Hohenberg–Kohn theorem states that the electron density alone is necessary to find the ground state energy of a system of N electrons, and that the energy is a unique functional of the density [2]. Unfortunately, the precise form of that functional is presently not known. However, we do have reasonably good approximations, although the Hohenberg–Kohn theorem does not offer a specific method to compute the electron density. The solution for a slow varying density is given by the Kohn–Sham formalism [3], where an auxiliary system of non-interacting

electrons in the effective potential is introduced, and the potential is chosen in such a way that the non-interacting system has exactly the same density as the system of interacting electrons in the ground state.

The Kohn–Sham (KS) equations below need to be solved iteratively until the self-consistent charge density is found:

$$\left[-\frac{1}{2}\nabla^2 + v_{\text{eff}}(r)\right]\varphi_i(r) = \varepsilon_i\varphi_i(r) \tag{7.5}$$

with the effective potential given by

$$v_{\text{eff}}(r) = v(r) + \int \frac{n(r')}{|r-r'|}\mathrm{d}r' + \frac{\delta E_{\text{xc}}[n]}{\delta n(r)} \tag{7.6}$$

where $v(r)$ is the external potential (e.g., due to ions) and $E_{\text{xc}}[n]$ is the exchange–correlation energy functional. The exact form of this functional is not known and has to be approximated. The density is given by

$$n(r) = \sum_{\text{occ}} |\varphi_i(r)|^2 \tag{7.7}$$

where the sum is over the N lowest occupied eigenstates. For the slowly varying density Kohn and Sham introduced the local density approximation (LDA):

$$E_{\text{xc}}[n] = \int \varepsilon_{\text{xc}}(n(r))n(r)\mathrm{d}r \tag{7.8}$$

where $\varepsilon_{\text{xc}}[n]$ is the exchange–correlation energy per particle of a uniform electron gas of density n. It is important to keep in mind that it is the electron density that is the "output" of the KS equations. Strictly speaking, the eigenvalues of the KS equations $\{\varepsilon_i\}$ have no direct physical meaning; nevertheless they are often very useful when the single particle electronic spectra (band structures) are discussed. The reasons behind the tremendous success of the Kohn–Sham theory are easy to identify. By solving essentially a single electron equation not much different from that due to Hartree, but including the effects of exchange and correlation, one gets an upper estimate of the ground state energy of a many-body system! The theory is variational, and thus forces acting on the atoms can be calculated. The equation, however, is non-linear and an iterative solution is needed.

Typically, the KS equations are projected onto a particular functional basis set, and the resulting matrix problem is solved. In terms of the basis, when solving KS equations one has two options. It is possible to discretize the equations in real space (this amounts to using δ-functions as a basis set) and solve them directly; these are so-called real space techniques [4]. Alternatively, one can choose a

complete set of conventional functions. There are two major functional basis set types presently employed. For periodic systems plane waves offer an excellent expansion set which along with the fast Fourier transformations affords an easy-to-program computational scheme, the accuracy of which can be systematically improved by increasing the number of plane waves [5]. For systems with strong, localized potentials such as those of the first row elements, a large number of plane waves are necessary in the expansion, and calculations require the use of ultra-soft pseudopotentials (see below) to be feasible. The second choice is to use local orbitals such as atomic orbitals or any other spatially localized functions. Among the advantages of a localized basis set are a smaller number of basis functions and sparsity of the resulting matrix due to the orbital's short range. The disadvantages are the complexity of multi-center integrals one needs and the absence of the systematic succession of approximations, since the set is typically either under-complete or over-complete. In both cases calculations are computer intensive.

7.2.3 Pseudopotential

Most likely the DFT-LDA approach would have been limited to small molecules if it were not for a pseudopotential method. Since only the valence electrons are involved in bonding and these electrons see a weaker potential due to screening by the core electrons, one can substitute the full Coulomb potential due to ions $v(r)$ with a smooth pseudopotential. This effectively reduces the number of electrons one needs to consider to the valence electrons only. For example, only 4 and not 14 electrons are needed for Si! The practical importance of this approximation should not be overlooked, a typical diagonalization algorithm scaled as N^3 with the size of the matrix, thus for silicon we get a factor of 42 for the speed-up! The most straightforward way to introduce a pseudopotential is due to Philips and Kleinman [6]. Today pseudopotentials used in electronic structure calculations may be broadly divided into three classes: the hard norm-conserving pseudopotentials [7], soft pseudopotentials [8] and Vanderbilt-type ultra-soft pseudopotentials [9]. The "softness" refers to how rapidly the potential changes in real space. The analogy comes from expanding a step function in a Fourier series; it takes a large number of plane waves to eliminate spurious oscillations at the step edge. On the other hand a "softer" function such as hyperbolic tangent can be expanded with greater ease. In general, hard pseudopotentials are more transferable. The choice of pseudopotential is in part dictated by the choice of a basis set used in the calculation. The use of local orbitals allows for a much harder pseudopotential. We will return to this point when discussing supercells.

7.2.4 Energy Minimization and Molecular Dynamics

Once the solution of KS equations is found, the total energy in the LDA is given by

$$E_{\text{total}} \approx \sum_i \varepsilon_i - \frac{1}{2}\iint \frac{n(r)n(r')}{|r-r'|}\mathrm{d}r\,\mathrm{d}r' + \int n(r)\{\varepsilon_{\text{xc}}(n(r)) - \mu_{\text{xc}}(n(r))\}\mathrm{d}r \quad (7.9)$$

where the exchange–correlation potential is given by $\mu_{\text{xc}} \equiv \frac{\mathrm{d}}{\mathrm{d}n}\{\varepsilon_{\text{xc}}(n(r))n(r)\}$. Now all ground state properties of the system can in principle be calculated. In particular, since we are using the Born–Oppenheimer approximation, the total energy of the electronic system, which is a function of the ionic positions $\{\vec{R}_1,\ldots,\vec{R}_i,\ldots,\vec{R}_N\}$, can be used as an inter-atomic potential. Note that unlike potential functions used in classical molecular dynamics or molecular mechanics methods, the energy function $E_{\text{total}}(\vec{R}_1,\ldots,\vec{R}_N)$ is not a sum of pair-wise interactions $\frac{1}{2}\sum_{i,j}V_{i,j}$ but a true many-body interaction energy computed quantum mechanically! One can easily calculate a force acting on any atom i in the direction α using the so-called Hellman–Feynman theorem $\left(\frac{\partial E}{\partial \lambda} = \left\langle \varphi(\lambda)\right| \frac{\partial H}{\partial \lambda}\left|\varphi(\lambda)\right\rangle\right)$ which is a rediscovery of the Ehrenfest result:

$$F_i^{\alpha} = \frac{\partial E_{\text{total}}}{\partial R_i^{\alpha}},\ \alpha = x, y, z \quad (7.10)$$

At this point one can either find the lowest-energy atomic configuration by employing an energy minimization technique such as damped molecular dynamics or a conjugate gradient method. Alternatively, a real molecular dynamics (MD) simulation can be launched. One has to keep in mind, however, that electronic frequencies $\frac{E_i - E_j}{\hbar}$ are much higher than a typical phonon frequency ω and for a stable simulation the time step needs to be a small fraction of the characteristic atomic period. The calculation then proceeds as follows. The KS energy is first calculated in a self-consistent manner for the initial atomic configuration, the Hellman–Feynman forces are evaluated and atoms are moved to the next time step via some MD algorithm (Verlet, Gear, etc. [10]). At the new configuration the KS equations are solved again, and the procedure is repeated. Needless to say, these are very expensive calculations. They offer a significant advantage if a temperature dependence of a particular quantity is sought, since MD can be performed at finite temperature. For example, the Fourier transform of the velocity auto-correlation function gives the vibration spectrum, thus calculations performed at different temperatures would give the temperature dependence of the phonon frequency.

7.2.5 Supercell/Slab Technique

As we have mentioned before the plane wave method is particularly well suited for studying periodic systems. However, many systems of interest, and particularly interfaces and surfaces, are manifestly non-periodic! Thus an artificial

system with periodicity is created to simulate them. The geometry is often referred to as slab or supercell. We shall illustrate the idea for the case of a surface. Here one clearly deals with a system in which the periodicity in one direction (that perpendicular to the surface) is broken. To perform surface calculations with a plane wave basis set a large simulation cell or a supercell is introduced in order to maintain artificial periodicity. A supercell contains a slab of bulk material (with many unit cells of the corresponding crystal) and vacuum slab in the direction perpendicular to the surface as illustrated in Fig. 7.2 for the (101) surface of PtSi. Si (Pt) atoms are represented with yellow (blue) color. The [101] direction is along the long side of the supercell. In the two directions parallel to the surface the supercell has the usual bulk dimensions, and the periodic boundary conditions are used without any change. The periodic boundary condition in the direction normal to the surface is applied for the supercell dimension, rather than the physical crystal cell side. Thus the "universe" is filled with infinite parallel slabs of PtSi of certain thickness, separated by infinite parallel slabs of vacuum. It is crucial that the length of a supercell in the direction normal to the surface is large enough to eliminate any spurious interactions between the cells across the vacuum region. The thickness of a slab should be sufficient for bulk properties to be restored in the middle of it. The supercell obviously creates two surfaces, and it is advisable to use a symmetric termination of the slab.

In principle, the larger the supercell is chosen the better it approximates true surface (or rather a set of two identical surfaces). However, the calculation also becomes more demanding, as we shall now demonstrate. In the case of a periodic system we write the eigenfunctions $\psi_{n,k}(r)$ of the KS equations as Bloch functions:

$$\psi_{n,k}(r) = u_{n,k}(r)\mathrm{e}^{\mathrm{i}kr} \tag{7.11}$$

where $u_{n,\mathbf{k}}(r)$ is a lattice periodic function, n is the band index and a wave vector $\mathbf{k}$ belongs to the first Brillouin zone (BZ). Since $u_{n\mathbf{k}}(\mathbf{r})$ is periodic, it can be expanded over the reciprocal lattice:

$$u_{n,k}(r) = \sum_{G'} \varphi_{n,k}(G')\mathrm{e}^{\mathrm{i}\vec{G}'\vec{r}} \tag{7.12}$$

where $\mathbf{G}'$ are the reciprocal lattice vectors. This expansion goes to infinity! Note that we actually deal with two types of infinities here. One is due to the infinite

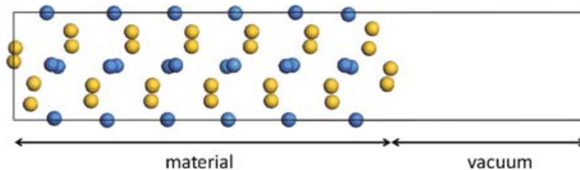


Fig. 7.2 Supercell used to simulate the (101) surface of PtSi. Si (Pt) atoms are represented with *yellow* (*blue*) color. The [101] direction is along the long side of the supercell (*See* Color Insert)

periodic nature of the crystal and is captured by the wave vector **k**; the other comes from this expansion. For practical purposes the sum over $\mathbf{G}'$ is restricted to plane waves with kinetic energy below a given cutoff energy E_{cut}. Thus, defining the set $\Omega(\mathbf{G})$

$$\Omega(G) := \left\{ \frac{\hbar^2}{2m} \left| \vec{k} + \vec{G} \right|^2 \leq E_{cut} \right\} \tag{7.13}$$

we obtain the following expansion of the Kohn–Sham wave functions:

$$\psi_{n,\vec{k}}(r) = \sum_{G \in \Omega(G)} \varphi_{n,k}(G) e^{i(\vec{G}+\vec{k})\vec{r}} \tag{7.14}$$

The cutoff energy E_{cut} controls the numerical convergence and depends strongly on the elements which are present in the system under investigation. For example, first row elements with strong potentials require higher cutoff energy. Here we immediately see the weakness of the supercell method. In the direction normal to the surface, the reciprocal cell vectors $|\vec{G}_{\perp}|$ are very short due to a large length of the direct space cell (often many multiples of the physical cell lattice constant). Thus a very large number of plane waves are needed to reach the convergence. This is the price one has to pay for the artificial periodicity. The introduction of ultra-soft pseudopotentials made these calculations practical. The localized basis set would still have the advantage of being insensitive to the simulation cell size; however, the range of the orbitals should be sufficient to describe the vacuum decay.

7.2.6 Calculating Band Alignment and Dielectric Constants

Among the most useful applications of the DFT-LDA scheme, from the gate dielectric development point of view, are calculations of the band discontinuity at the interface and of the dielectric constant. The discontinuity can be estimated using the reference potential method originally introduced by Kleinman [11]. Van de Walle and Matin proposed using the macroscopically averaged electrostatic potential as reference energy [12]. The method requires calculating a heterojunction AB in either slab (in this case you would have free surfaces) or supercell geometry to compute the average reference potential across the interface and two additional bulk calculations to locate the valence band top (VBT) in materials A and B with respect to the average potential. For a supercell (or a slab) containing the interface one calculates the average potential using the formula:

$$\bar{V}(z) = \frac{1}{d_1 d_2} \int_{z-d_1/2}^{z+d_1/2} \mathrm{d}z' \int_{z'-d_2/2}^{z'+d_2/2} \mathrm{d}z'' \, V(z'') \, . \tag{7.15}$$

where $V(z)$ is obtained by the xy-plane averaging (a simple $\frac{1}{(a_x \cdot a_y)} \iint_{\text{cell}} dx\, dy$ integration) of the electrostatic potential:

$$V(r) = -\sum_i \frac{Z_i e^2}{|r - R_i|} + e^2 \int \frac{n(r')}{|r - r'|} \mathrm{d}r' \tag{7.16}$$

The parameters d_1 and d_2 are the inter-planar distances along the z direction (normal to the interface) in materials A and B, respectively. This produces a smooth reference potential. Assuming that far away from the interface the potential reaches its bulk value one can place corresponding VBTs with respect to the average potential on both sides of the interface using the bulk reference, and thus determine the VBO. The conduction band offset has to be inferred using the experimental values of the band gaps, since those are seriously underestimated in the DFT-LDA calculations.

Calculating the dielectric constant is less straightforward due to the periodic boundary conditions used in most first principles codes. In brief, it is the absence of the surface in an infinite periodic solid that causes the problem. Vanderbilt has shown that the change in electronic polarization can be calculated using the geometric or Berry phase of electrons [13]:

$$P_\alpha^{\text{el}} = \frac{i}{\Omega} \sum_{ki} \left\langle u_{ki} \middle| \frac{\partial}{\partial k_\alpha} \middle| u_{ki} \right\rangle \tag{7.17}$$

where Ω is the unit cell volume, k is the Bloch vector and u_{ki} is the cell periodic part of the Bloch wave function. Once the change in polarization with respect to a reference state of the system is determined, Born effective charges Z_{ia}^{*M} can be evaluated, and the dielectric constant is given by

$$\varepsilon_{\alpha\beta} = \varepsilon_{\alpha\beta}^{\infty} + \frac{4}{\pi} \sum_i \frac{Z_{i\alpha}^{*M} Z_{i\beta}^{*M}}{\omega_i^2 - \omega^2} \tag{7.18}$$

The electronic contribution $\varepsilon_{\alpha\beta}^{\infty}$ can be computed using the linear response theory. The values thus computed typically overestimate the experiment by about 20%, mainly due to the error in the band gap. A semi-empirical “scissor” correction is then used in which the conduction bands are moved up in energy by hand to match the experimental spectrum.

7.2.7 Ab Initio Packages

Today many first principles codes are available. An example of a real space code is PARSEC [4]. VASP [14] and CASTEP [15] are plane wave codes. FIREBALL [16], SIESTA [17] and DMol [18] are local atomic orbital codes. The work horse of computational chemistry GAUSSIAN is a local orbital code

using atomic orbitals expanded in terms of Gaussians to simplify multi-center integrations [19]. Linear response calculations can be performed with PWSCF [20] and *Abinit* [21]. Overall, DFT-LDA calculations give very accurate ground state properties such as structural parameters, elastic constants and relative energies of different phases. The most serious drawback of the theory is its inability to describe the excited states, and thus to predict a band gap. Several methods have been developed to address this problem, such as the exact exchange method [22], GW method [23] and Bethe–Salpeter method [24]. Unfortunately, all of these techniques require a significant increase in computational time. To learn more about the applications of the DFT-LDA formalism to high-*k* dielectrics we refer the reader to reference [25].

7.2.8 Beyond the DFT-LDA

Despite its astounding success in materials theory, the failures of the DFT-LDA scheme are numerous, systematic and well documented [26]. Many of these failures occur in transition metal oxides where the LDA, being a mean field theory, fails to properly account for electron correlations (strictly speaking, it is not possible to separate exchange and correlation in the LDA-DFT formalism). The physical reason for this failure is a relatively high degree of electron localization in the TM d-shells. Perdew and Zunger have shown that the self-interactions result in significant errors in single particle energy levels [27]. Self-interaction corrections (SIC) have been successfully implemented and used for calculations of TM oxides [28]. Most recently, a very attractive scheme avoiding orbital-depending potentials was suggested by Filippetti and Spaldin [29]. Another way to at least partially account for the electron correlation is the so-called LDA+U method [30]. Lee and Picket have successfully used it to describe magnetic ordering in Sr_2CoO_4 [31].

7.3 A Brief Overview of Recent Theoretical Results

Many of the high-*k* dielectrics also happen to be important ceramic-forming materials. Hafnia and zirconia are no exception, and as such they are relatively well studied. However, the type of questions one would ask about electronic materials is very different from that commonly asked about ceramics. Thus the electronic properties of hafnia are not as well characterized. One of the first theoretical studies of the structural and electronic properties of different phases of high-*k* materials in general has been by Medvedeva and co-workers who systematically studied the subgroup IVa transition metal dioxides using the linear muffin-tin orbitals in the atomic sphere approximation (LMTO-ASA) method [32]. In particular, for the cubic fluorite phase of HfO_2 they found the lattice constant and cohesive energy in good agreement with experiment. More

recently, the structural properties of HfO_2 and ZrO_2 were investigated by Lowther et al. using the ab initio pseudopotential plane wave method [33]. They reported elastic constants and relative stability of the high-pressure phases and showed similarities between ZrO_2 and HfO_2. Other first principles studies of electronic, structural and vibrational properties of zirconia and hafnia, including the high-temperature phases, have been reported [34, 35, 36, 37, 38, 39]. In particular, Vanderbilt's group reported the first theoretical study of bulk amorphous zirconia [39]. Overall theoretical results show reasonable agreement with each other and with experiment.

The static dielectric constants of hafnia and zirconia have also been computed and found to be highly dependent on the crystal phase [36]. They are also highly anisotropic for low-symmetry phases, with an especially large dielectric response in the basal plane of the tetragonal structure. The large dielectric constants arise from (i) the presence of relatively low-frequency polar phonon modes and (ii) anomalously large Born effective charges that result from the hybridization between the O p- and metal d-states. Rignanese found that the tetragonal phase has the largest and most anisotropic dielectric constant [38], in qualitative agreement with the earlier result by Vanderbilt [36].

First principles methods can also be used to study phase transitions. A powerful technique is to combine a model Hamiltonian based on first principles calculations with Monte Carlo simulations [40, 41, 42, 43]. In the case of purely displacive phase transitions, a soft mode (a phonon mode with a frequency that falls to zero at the transition temperature) can be identified using first principles lattice dynamics simulations [44, 45, 46, 47, 48]. The cubic-to-tetragonal transition in zirconia has recently been studied using ab initio molecular dynamics [49].

Electrically active point defects in hafnia can act as electron or hole traps and are believed to play a significant role in the negative bias temperature instability (NBTI) [50, 51, 52]. It is generally believed these are oxygen-related defects, in other words oxygen vacancies or oxygen interstitials. A comprehensive theoretical study of both types of defects was done by Foster and co-workers using DFT [52]. They considered oxygen incorporation into hafnia from molecular and atomic oxygen and found that atomic incorporation is energetically most favorable. They also studied charged defects including charge transfer reactions between the defects. Whether a defect is a charge trap depends on the position of the defect-related states with respect to the valence band top of hafnia as well as the band alignment with silicon/silica and metal. The local density approximation to DFT employed by Foster typically underestimates the band gap, and the authors "corrected" the defect energy levels using experimental value for the band gap. More recently, Xiong and Robertson have calculated defect levels using the screened exchange method which gives a reasonably good value of 5.75 eV for the band gap of monoclinic hafnia [53]. In addition, oxygen vacancy levels in ZrO_2 were calculated using the GW approximation (essentially the first-order many-body correction to the self-energy operator starting from the LDA result, G stands for Green's function and W for the screened Coulomb potential) by Kralik et al. [54]. An interfacial SiO_2 layer is always

present between silicon and hafnia, and substitutional defects such as silicon in hafnia or hafnium in silicon or silica are of interest. The comparative analysis of substitutional defects was performed by Scopel et al. [55]. They found that the defect formation energy strongly depends on the chemical environment. Under oxygen-rich conditions substitutional silicon in hafnia is the most likely defect. On the other hand, the formation of substitutional Hf defects in SiO_2 is less likely under oxygen-rich conditions than under hafnia-rich conditions.

Theoretical calculations of ZrO_2 surfaces have been first reported by Christensen and Carter [56]. They investigated many surface orientations for all three zirconia polymorphs. Recently, Mukhopadhyay and co-authors have reported a density functional study of monoclinic hafnia surfaces [57]. They have considered only stoichiometric terminations and identified (111) and $\left(11\bar{1}\right)$ surfaces as most stable.

After the early work on Si/SiO_2 interfaces [58, 59], several authors reported DFT-LDA calculations of interfaces of high-*k* dielectrics with metals and semiconductors [60, 61, 62, 63, 64]. Puthenkovilakam et al. considered the interfaces between the (001) surfaces of tetragonal zirconia (t-ZrO_2) or zircon ($ZrSiO_4$) and a silicon (100) substrate within the local density approximation [62]. They find that ZrO_2/Si interfaces exhibit partial occupation of zirconium dangling bonds (Zr d-states at the Fermi level) when the zirconium coordination is reduced from its bulk coordination. Hydrogen passivation of zirconium atoms, as well as oxygen bridging at the interface, can remove the partial occupancy of d-orbitals at the Fermi level. The calculated band offsets of these interfaces show asymmetric band alignments, with conduction band offsets between 0.64 and 1.02 eV and valence band offsets between 3.51 and 3.89 eV. By contrast, the $ZrSiO_4/Si$ interface provides a more symmetric band alignment, with a much higher conduction band offset of 2.10 eV and a valence band offset of 2.78 eV. These results suggest that $ZrSiO_4$ may form an excellent interface with silicon in terms of its electronic properties and therefore maybe a suitable candidate for replacing SiO_2 as a gate insulator in silicon-based field-effect transistors. Recently Dong and co-authors reported on the theory of the Schottky barrier formation at the $Ni/ZrO_2(001)$ interface [63]. Their suggestion of using a heterovalent metal interlayer to tune the barrier, though appearing to be impractical, illustrates the power of theory in investigating hypothetical systems before doing experiments.

Several density functional studies of electron transport in oxides have been reported. Fonseca and co-authors used a combination of first principles density functional theory and non-perturbative scattering theory to investigate the effect of point defects on the hole leakage current through ultra-thin hafnia films [64]. They found that the neutral bulk vacancies and an interface vacancy along the Si–O–Si bond have little impact on the leakage current. On the contrary, an interface vacancy along the Hf–O–Si bridge and an interstitial B atom in the HfO_2 region introduce states in the Si band gap, thus strongly enhancing the leakage current at a low bias. Recently, Evans and co-workers reported a transport study of the Si–SiO_2 system at the level of Boltzmann

equation using first principles derived potentials when calculating the collision integral [65]. This approach seems particularly useful, since the perturbing potential being a ground state property is well reproduced within the DFT-LDA formalism, while semi-classical transport equations allow consideration of macroscopic device structures.

7.4 Band Alignment at the Si/SiO_2, Si/HfO_2, SiO_2/HfO_2 and HfO_2/Mo Interfaces

When studying the gate stack of the MOS capacitor, one of the most intriguing questions is the overall line-up of electronic bands in various materials of which the stack is built. Since the processed gate stack has multiple interfaces (see Fig. 7.3) it seems natural to estimate the band alignment at each interface first and then build up the band diagram across the stack by adding the corresponding shifts. However, it is still not clear if the overall alignment of the stack can be reproduced from this piecewise approach. The question is currently under intense investigation. We start our discussion with the band alignment at the Si–SiO_2 interface. A simple estimate of the conduction band offset using the metal-induced gap state (MIGS) model is given by [66]

$$\phi = (\chi_a - \Phi_a) - (\chi_b - \Phi_b) + S(\Phi_a - \Phi_b) \tag{7.19}$$

Here χ is the electron affinity, Φ_i is the charge neutrality level of material i measured from the vacuum level, S is an empirical dielectric pinning parameter describing the screening by the interfacial states and subscripts a and b refer to Si and dielectric, respectively. If $S = 1$ the offset is given by a difference in electron affinities as was originally proposed by Schottky [67]. Alternatively,

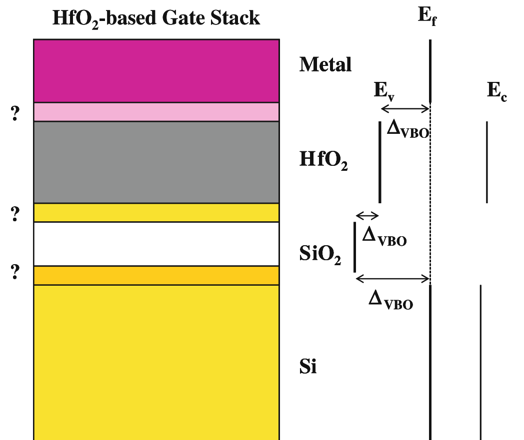


Fig. 7.3 Schematic of a multilayer gate stack based on HfO_2. A plausible band alignment across the stack is also shown (*See* Color Insert)

for $S=0$ we get the strong pinning or the Bardeen limit [68]. The pinning parameter can be estimated by the empirical formula [66]

$$S = \frac{1}{1 + 0.1(\varepsilon_\infty - 1)^2} \tag{7.20}$$

where ε_∞ is the high-frequency component of the dielectric constant. Electron affinities are typically known experimentally. Tersoff proposed a simple way to estimate the charge neutrality level position associating it with the branch point of the complex band structure of the dielectric [69]. For β-crystobalite the complex band structure gives charge neutrality level 5.1 eV (the value is actually rescaled using the ratio of the experimental and calculated band gaps) above the valence band maximum [70], thus placing it 4.8 eV below the vacuum if we assume an electron affinity χ of 0.9 eV. The imaginary wave vector along the *c*-axis of the tetragonal cell has a length of 1.3 Å^{-1} at the branch point. The electron affinity and charge neutrality level of Si with respect to vacuum are 4.0 and 4.9 eV, respectively. The pinning parameter S of SiO_2 is 0.9, thus the conduction band offset comes out as 3.1 eV in rather good agreement with experiment.

Insofar as the ab initio calculations are concerned, we have investigated the band alignment at the Si–SiO_2 interface using a local orbital variation of the reference potential method and found good agreement with experiment [58]. The salient feature of that work was building of the theoretical interface structure by explicitly modeling the oxidation reaction. Thus the interface showed both structural and chemical disorders and contained a very thin layer of sub-oxide. On the other hand, Tang et al. used a plane wave method for crystalline Si–SiO_2 interfaces and reported the valence band offset much smaller than experiment [71]. More recently, Tuttle et al. have also reported small valence band offsets computed with a plane wave method and approach similar to that of Tang and co-authors [72].

7.4.1 Si/HfO_2 Interface

The ab initio calculation of the valence band discontinuity Δ_{VBO} at the Si–HfO_2 interface using the reference potential method is illustrated in Fig. 7.4. Here we follow the discussion of reference 64. The planar averaged reference potential for the Si–HfO_2–Si slab is plotted along the direction normal to the interface. The average values in the HfO_2 and Si regions of the slab are –18.9 and –17.0 eV, respectively. From the corresponding bulk calculations we know that the distances from the reference potential to the valence band maximum Δ_{RV} are 12.0 and 13.0 eV in HfO_2 and Si, respectively. The resulting valence band discontinuity $\Delta_{VBO}=2.9$ eV, which is roughly between the value of 3.2 eV estimated by Robertson [73] and the experimental value of 2.2 eV [74].

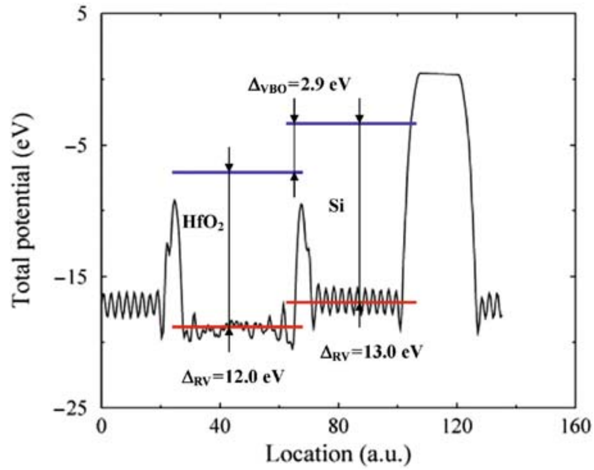


Fig. 7.4 The valence band offset between Si and HfO_2 is calculated using the reference potential method. The average reference potential is indicated with *red lines*, and the valence band maxima with *blue lines*. The discontinuity is estimated to be 2.9 eV (*See* Color Insert)

In Fig. 7.5 we show the complex band structure calculated for monoclinic HfO_2 [70]. The complex band structure is calculated by posing the following question: given the Hamiltonian what are the solutions corresponding to certain energy (a real number). Unlike the case of the usual band structure calculations when one finds the set of eigenvalues corresponding to a real wave vector, here one does not require the wave vector to be real. For an infinite periodic solid even a well-behaved exponentially decaying solution would be inconsistent with the boundary conditions. However, in the case of the surface or, for that matter, for any kind of a symmetry-breaking defect these are perfectly legitimate solutions. For HfO_2 the LDA band gap is calculated to be 3.5 eV, therefore a rescaled value of the charge neutrality level is estimated to be

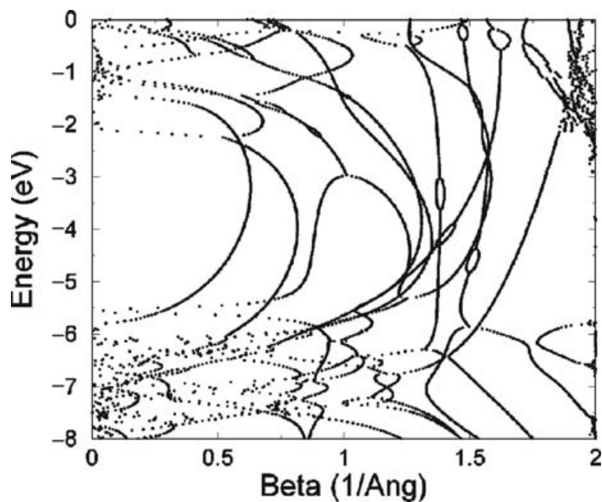


Fig. 7.5 The complex band structure of m–HfO_2 in the near gap region. The charge neutrality level is 2.3 eV above the band valence band top as calculated. The band gap is calculated to be 3.5 eV, therefore the rescaled value of the charge neutrality level is estimated to be 3.8 eV

3.8 eV using an experimental band gap for m–HfO_2 of 5.8 eV. Assuming the commonly used value of electron affinity of 2.5 eV, the charge neutrality level is 4.5 eV with respect to vacuum. The length of the imaginary wave vector at the branch point along the (001) direction is $\sim$0.3 Å^{-1}. It is interesting to note that evanescent states penetrate much deeper in m–HfO_2 (about 3.3 Å) than in SiO_2 (about 1.5 Å). This together with a higher ε_∞ makes m–HfO_2 a strongly pinning material (indeed, $S = 0.53$). The complex band is relatively flat in the vicinity of the branch point. This suggests a relative "insensitivity" of the result. The conduction band offset calculated using MIGS equation (7.19) above is only 1.4 eV! The estimate can be improved by assuming that highly doped n-type Si behaves almost as a metal, and using the metal/insulator formula for a Schottky barrier,

$$\phi = S(\Phi_m - \Phi_b) + (\Phi_b - \chi) \tag{7.21}$$

Here Φ_m is the work function of Si. The resulting conduction offset is 1.8 eV which is in much better agreement with the DFT result and experiment.

7.4.2 SiO_2/HfO_2 Interface

Hafnia can be deposited on a Si wafer by several techniques: atomic layer deposition (ALD), metalorganic chemical vapor deposition (MOCVD) or physical vapor deposition (PVD), using various precursors [75]. However, in all cases, a thin SiO_2 layer, grown either intentionally or spontaneously, is present at the interface between the high-*k* film and Si substrate after the standard fabrication processing is completed. The band offset between SiO_2 and HfO_2 is unknown but clearly determines the overall alignment of the gate stack. It is possible that our failure to correctly include the dipole layer at the oxide–oxide interface contributes to our inability to explain many experimental results in these advanced gate stacks [76]. We have constructed several atomistic SiO_2/HfO_2 models which differ by the interfacial oxygen coordination, HfO_2 phases and strain and studied the interface using density functional theory [77]. In every structure the transition from one oxide to another is achieved via a Si–O–Hf bridge bond. This ensures a clear band gap free of defect states. We use these models to calculate the band discontinuity, thus relating the microscopic structure of the stack to its electric properties. The analysis of trends thus computed allows us to put forward the following description of the band alignment. The valence band offset is found to vary between –2.0 and 1.0 eV depending on the microscopic structure of the interface and to depend strongly on the average coordination of the interface oxygen. The Schottky limit value of 1.6 eV is recovered for the fully oxidized interface. We suggest that the final band offset value is mostly determined by the interface layer polarizability, which in turn is directly related to the average coordination of the interface oxygen.

7.4.3 HfO_2/Mo Interface

One of the most significant challenges to using hafnia as a gate dielectric in CMOS technology is the absence of a suitable p-type gate metal [76]. That is a metal with such a *Schottky barrier* to HfO_2 that aligns its Fermi level with the top of the valence band of Si on the other side of the metal oxide semiconductor (MOS) capacitor. The misalignment of these two energies results in a larger threshold voltage and thus reduces the drive current available at a given bias which is the principal measure of the device performance [78]. Even in the case of a "simple" poly-Si gate, Si-metal bonds at the metal/oxide interface and the ensuing Fermi level pinning have been suggested as a possible explanation of the band misalignment [79]. Another possible reason could be point defects such as oxygen vacancies. We have shown that the failure to find an appropriate metal is rooted in our failure to understand the fundamental physics of the transition metal oxide/metal interface formation [80]. It is the Schottky barrier that is the critical parameter here. However, the *work function* is a crucial component of the band alignment and has to be about 5.0 eV with respect to vacuum to be close in energy to the valence band of Si. There are only a handful of pure metals with work functions that large, i.e., W, Mo, Pd, Pt, Os, Re, Ru, Rh, Au, Co and Ni [81]. Co and Ni are fast diffusers and would not be the first choice, Au is difficult to etch, but W, Mo, Pd, Os, Re, Ru, Rh and Pt are potential contenders. Using density functional theory to build an atomistic model of the Mo/HfO_2 interface we calculate the Schottky barrier of 2.8 eV in near-perfect p-type band alignment. The plane average electrostatic potential across the Mo/HfO_2 interface is shown in Fig. 7.6. The valence band of the oxide is 3 eV below the Fermi level, which translates into a perfect p-type alignment with Si. However, when we investigate the thermodynamic stability of the interface and the corresponding alignment with respect to a metal–dielectric oxygen exchange reaction, we find that both are unstable! We discover a low-energy interfacial defect, which we call the extended Frenkel pair. It forms via transfer of oxygen across the interface resulting in a vacancy in the oxide and an interstitial in the metal. This defect causes an almost half a volt change in the Schottky barrier height! This behavior is expected of most large work function metals in contact with a transitional metal oxide, with the exception of Ru, Rh and Os. The vacancy formation energy is lowered dramatically due to the large oxidation enthalpy of the metal and the availability of the electronic reservoir (the Fermi sea of the metal) to accommodate electrons associated with a neutral vacancy. The presence of a high work function metal with a large oxide formation enthalpy significantly alters the defect chemistry at the metal/high-*k* interface. It effectively lowers the oxygen vacancy formation energy which results in the intrinsic instability of the interface dipole. This understanding helps us to identify better gate metal candidates. For a p-type alignment among the eight pure metals the best choices would be Ru (RuO_2 $\Delta H = -75.1$ kcal/mol or -1.02 eV per oxygen), Rh (RhO_2 $\Delta H = -45.17$ kcal/mol or -0.98 eV per oxygen)

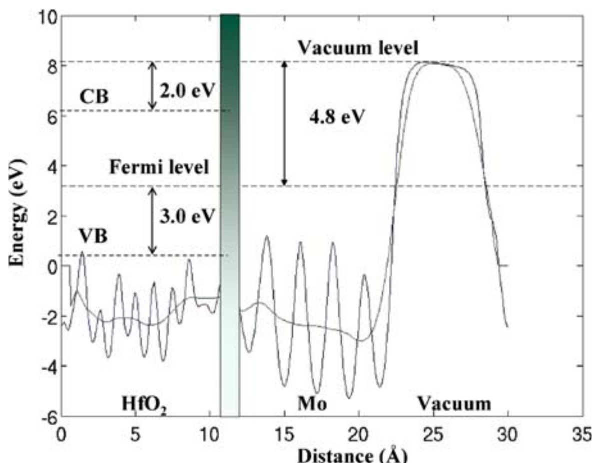


Fig. 7.6 The average reference potential across the Mo–HfO_2 heterojunction. The vacuum level, conduction and valence band of HfO_2 and the Fermi level are indicated (*See* Color Insert)

and Os (OsO_4 $\Delta H = -90.5$ kcal/mol or -0.98 eV per oxygen) with work functions of 4.71, 4.98 and 4.83 eV, respectively.

7.5 Conclusions

Density functional theory is being rather successfully used to support materials development for high-*k* dielectric gate stacks in advanced CMOS technology despite its limitations with respect to the excited states and computational expense associated with a large number of atoms in these systems. Many properties of bulk structures and thin films such as the atomic and electronic structure or linear dielectric response can be reliably calculated from first principles. However, the accuracy of many important parameters such as the band offset is still not sufficient for a direct engineering support. Nevertheless, the qualitative picture and trends predicted by first principles calculations are of great value and can be used to guide the experimental effort.

Acknowledgments I wish to thank many colleagues for insightful discussions we have had over the years and my graduate students at the University of Texas, Onise Sharia, Xuhui Luo and Jaekwang Lee, for their hard work and help with the manuscript. This work in part is supported by the National Science Foundation under grants DMR-0548182 and DMR-0606464 and by the Office of Naval Research under grant N000 14-06-1-0362.

References

1. M. Born and R. Oppenheimer, Ann. Phys. **84**, 458 (1927).
2. P. Hohenberg and W. Kohn, Phys. Rev. **136**, 864 (1964).
3. W. Kohn and L.J. Sham, Phys. Rev. **140**, 1133 (1965).

4. J.R. Chelikowsky, N. Troullier, and Y. Saad, Phys. Rev. Lett. **72**, 1240 (1994).
5. M.C. Payne, M.P. Teter, D.C. Alan, T.A. Arias, and J.D. Joannopoulos, Rev. Mod. Phys. **64**, 1045 (1992).
6. J.C. Philips and L. Kleinman, Phys. Rev. **116**, 287 (1959).
7. D. Hamann, M. Schluter, Chiang, Phys. Rev. Lett. **43**, 1494 (1979).
8. N. Trulier, and J.L. Martins, Phys. Rev. B **43**, 1993 (1991).
9. D. Vanderbilt, Phys. Rev. B **41**, 7892 (1990).
10. M.P. Allen and D.J. Tildesley, *Computer Simulation of Liquids*, (Clarendon Press, New York, 1988).
11. D.M. Bylander and L. Kleinman, Phys. Rev. B **36**, 3229 (1987).
12. C.G. Van de Walle and R.M. Martin, Phys. Rev. B **39**, 1871 (1989).
13. R.D. King-Smith and D. Vanderbilt, Phys. Rev. B **47**, 1651 (1993).
14. G. Kresse and J. Furtmuller, Phys. Rev. B **54**, 11169 (1996).
15. V. Milman, B. Winkler, J.A. White, C.J. Pickard, M.C. Payne, E.V. Akhmatskaya, and R.H. Nobes, J. Quant. Chem. **77**, 895 (2000).
16. J.P. Lewis, K.R. Glaesemann, G.A. Voth, J. Fritsch, A.A. Demkov, J. Ortega, and O.F. Sankey, Phys. Rev. B **64**, 195103 (2001).
17. J.M. Soler, E. Artacho, J.D. Gale, A. García, J. Junquera, P. Ordejón, and D. Sánchez-Portal, J. Phys.: Condens. Matter**14,** 2745 (2002).
18. B. Delley, J. Chem. Phys. **113**, 7756 (2000).
19. M.J. Frisch et al., GAUSSIAN 98, (Gaussian, Inc., Pittsburgh, PA, 1998).
20. S. Baroni, A. Dal Corso, S. de Gironcoli, P. Giannozzi, Plane wave self-consistent field (URL: http://www.pwscf.org).
21. X. Gonze, D.C. Allan, M.P. Teter, Phys. Rev. Lett. **68**, 3603 (1992). (URL: http://www.abinit.org.).
22. M. Städele, J.A. Majewski, P. Vogl, and A. Görling, Phys. Rev. Lett. **79**, 2089 (1997); M. Städele, M. Moukara, J. A. Majewski, and P. Vogl, Phys. Rev. B **59**, 10031 (1999).
23. F. Aryasetiawan and O. Gunnarsson, Phys. Rev. Lett. **74**, 3221 (1995).
24. M. Rohlfing and S.G. Louie, Phys. Rev. B **62**, 4927 (2000).
25. A.A. Demkov and A. Navrotsky, Eds., *Materials Fundamentals of Gate Dielectrics*, (Springer, Dordrecht, 2005).
26. R.O. Jones and O. Gunnarson, Rev. Mod. Phys. **61**, 689 (1989).
27. J.P. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).
28. M. Arai and T. Fujiwara, Phys. Rev. B **51**, 1477 (1995).
29. A. Filippetti and N.A. Spaldin, Phys. Rev. B **67**, 125109 (2003).
30. V.I. Anisimov and P. Kuiper, et al., Phys. Rev. B **50**, 8257(1994).
31. K.W. Lee and W.E. Pickett, Phys. Rev. B **73**, 174428 (2006).
32. N.I. Medvedeva, V.P. Zhukov, M. Ya Khodos, and V.A. Gubanov, Phys. Status Solidi B **160**, 517 (1990).
33. J.E. Lowther, J.K. Dewhurst, J.M. Leger, et al., Phys. Rev. B **60**, 14485 (1999).
34 A.A. Demkov, Phys. Stat. Sol. B **226**, 57 (2001).
35. X.Y. Zhao and D. Vanderbilt, Phys. Rev. B **65**, 075105 (2002).
36. X.Y. Zhao and D. Vanderbilt, Phys. Rev. B **65**, 233106 (2002).
37. R. Terki, H. Feraoun, G. Bertrand, et al., Comp. Mat. Sci. **33**, 44 (2005).
38. G.M. Rignanese, X. Gonze, G. Jun, et al., Phys. Rev. B **69**, 184301 (2004).
39. D. Vanderbilt, X.Y. Zhao, and D. Ceresoli, Thin. Solid Films **486**, 125 (2005); X.Y. Zhao, D. Ceresoli, and D. Vanderbilt, Phys. Rev. B **71**, 085107 (2005).
40. R.D. King-Smith and D. Vanderbilt, Phys. Rev. B **49**, 5828 (1994).
41. K. Parlinski, Z.Q. Li, et al., Phase Transitions **67**, 681 (1999).
42. D. Vanderbilt and W. Zhong, Ferroelectrics **206**, 181 (1998).
43. W. Zhong, D. Vanderbilt, and K.M. Rabe, Phys. Rev. B **52**, 6301 (1995).
44. P. Ghosez, E. Cockayne, U.V. Waghmare, et al., Phys. Rev. B **60**, 836 (1999).
45. M.T. Dove, Amer. Mineral. **82**, 213 (1997).

46. M. Sternik and K. Parlinski, J. Chem. Phys. **123**, 204708 (2005).
47. K. Parlinski, Z.Q. Li, and Y. Kawazoe, Phys. Rev. Lett. **78**, 4063 (1997).
48. A.P. Mirgorodsky, M.B. Smirnov, T. Merle-Mejean, et al., J. Mat. Sci. **34**, 4845 (1999).
49. S. Fabris, A.T. Paxton, and M.W. Finnis, Phys. Rev. B **63**, 094101 (2001).
50. S. Zafar, B.H. Lee, and J.H. Stathis, IEEE Electron Device Lett. **25**, 153 (2004).
51. M. Houssa, M. Aoulaice, S. Van Elshocht, S De Gent, G. Groeseneken, and M.M. Heyns, Appl. Phys. Lett. **86**, 173509 (2005).
52. A.S. Foster, F. Lopes Gejo, A.L. Shluger, and R.M. Neiminen, Phys. Rev. B **65**, 174117 (2002).
53. K. Xiong, J. Robertson, Microelectronics Engineering **80**, 408 (2005).
54. B. Kralik, E.K. Chang, and S.G. Louie, Phys. Rev. B. **57**, 7027 (1998).
55. W.L. Scopel, A.J.R. da Silva, W. Orellana, and A. Fazzio, Appl. Phys. Lett. **84**, 1492 (2004).
56. A. Christensen and E. Carter, Phys. Rev. B **58**, 8050 (1998).
57. A.B. Mukhopadhyay, J.F. Sanz, and C.B. Musgrave, Phys. Rev. B **73**, 115330 (2006).
58. A.A. Demkov and O.F. Sankey, Phys. Rev. Lett. **83**, 2038 (1999).
59. J. Neaton, D. Muller, and N. Ashcroft, Phys. Rev. Lett. **85**, 1298 (2000).
60. V. Fiorentini and G. Gulleri, Phys. Rev. Lett. **89**, 266101 (2002).
61. J. Robertson, J. Non-Crystalline Solids **303**, 94–100 (2002).
62. R. Puthenkovilakam, E.A. Carter, and J.P. Chang, Phys. Rev. B **69**, 155329 (2004).
63. Y.F. Dong, S.J. Wang, Y.P. Feng, and A.C.H. Huan, Phys. Rev. B **73**, 045302 (2006).
64. L.R.C. Fonseca, A.A. Demkov and A. Knizhnik, Phys. Stat. Sol. B **239**, 48 (2003).
65. M. Evans, X. Zhang, J. Joannopoulos, and S. Pantelides, Phys. Rev. Lett. **95**, 106802 (2005).
66. J. Robertson and C.W. Chen, Appl. Phys. Lett. **74**, 1168 (1999)
67. W. Schottky, Zeits. Physik **118**, 539 (1942).
68. J. Bardeen, Phys. Rev. **71**, 717 (1947).
69. J. Tersoff, Phys. Rev. B **30**, 4874 (1984).
70. A.A. Demkov, L. Fonseca, E. Verret, J. Tomfohr, and O.F. Sankey, Phys. Rev. B **71**, 195306 (2005).
71. S. Tang, R.M. Wallace, A. Sebaugh, and D. King-Smith, Appl. Surf. Sci. **135**, 137 (1998).
72. B. Tuttle, Phys. Rev. B **67**, 155324 (2003).
73. J. Robertson, J. Vac. Sci. Technol. B **18**, 1785 (2000).
74. Y.T. Hou, M.F. Li, H.Y. Yu, and K.L. Kwong, *Proceedings of the 2003 Symposia on VLSI Technology and Circuits* (VLSI 2003).
75. M. Ritala, M. Leskela, L. Niinisto, T. Prohaska, G. Friedbacher, and M. Grasserbauer, *Thin Solid Films* **250,** 72 (1994); P. Kisch et al., J. Appl. Phys. (2006).
76. R. Chau, IEEE Elec. Dev. Lett. **25**, 408 (2004).
77. O. Sharia, A.A. Demkov, G. Bersuker, and B.-H. Lee, Phys. Rev. B. **75**, 035306 (2007).
78. S.M. Sze, *Physics of Semiconductor Devices*, (Wiley, New York, 1981).
79. C.C. Hobbs, L.R.C. Fonseca, A. Knizhnik, V. Dhandapani, S.B. Samavedam, W.J. Taylor, J.M. Grant, L.G. Dip, D.H. Triyoso, R.I. Hegde, D.C. Gilmer, R. Garcia, D. Roan, M.L. Lovejoy, R.S. Rai, E.A. Hebert, H.-H. Tseng, S.G.H. Anderson, B.E. White, and P.J. Tobin, IEEE Trans. Elec. Dev. **51**, 971 and 978 (2004).
80. A.A. Demkov, Phys. Rev. B **74,** 085310 (2206).
81. H.B. Michaelson, J. Appl. Phys. **48**, 4729 (1977).

Chapter 8
Trapping Phenomena in Nanocrystalline Semiconductors

Magdalena Lidia Ciurea

Abstract In this chapter, trapping phenomena in nanocrystalline semiconductors (materials and devices) are presented and analyzed. The small number of atoms in a nanocrystalline semiconductor makes the contributions of the traps to different phenomena much more important as compared to a bulk semiconductor. The conventional (experimental) methods most frequently used for the investigation of traps are described. I also discuss which methods are suitable to be used for the trap investigation in nanocrystalline semiconductors and what are the trap parameters that can thus be obtained. The application of these methods, together with different non-conventional methods, to the study of the traps in nanocrystalline semiconductors, is presented. The role of the traps in possible applications as well as functioning problems of different devices is outlined.

8.1 Introduction

It is important to study the trapping phenomena in semiconductor materials and devices because these phenomena contribute to different properties of the materials and can modify some parameters of the devices. In nanocrystalline semiconductors, the traps play an even more important role. Indeed, a single trap in a nanocrystal of 1000 atoms represents a trap concentration much greater than any value attained in bulk semiconductors.

In the following, the definitions of a trap center and of the related parameters are presented. The differences between trapping and recombination centers are discussed. Next, the particularities of the traps in nanocrystalline semiconductors are described, together with the special phenomena that take place at the nanometer scale.

In Section 8.2, the main investigation methods of the traps are briefly presented. I focus on how to use these methods, rather than on giving full

M.L. Ciurea
National Institute of Materials Physics, Bucharest, Romania
e-mail: ciurea@infim.ro

A. Korkin, F. Rosei (eds.), *Nanoelectronics and Photonics*,

DOI: 10.1007/978-0-387-76499-3_8, © Springer Science+Business Media, LLC 2008

mathematical analyses on how these methods work. Because some of these methods raise difficulties when they are applied to nanocrystalline semiconductors, different non-conventional methods to investigate the traps in particular structures and devices were developed. Usually such non-conventional methods cannot give all the parameters that can completely characterize the traps, so they must be coupled among themselves or with some "classical" method whenever possible.

Section 8.3 contains the experimental results and possible applications obtained on the most used nanocrystalline materials by either conventional or non-conventional methods. The non-conventional methods are briefly described, to arouse the reader's interest. Section 8.4 summarizes the chapter.

It is well known that non-equilibrium free carriers (electrons and holes) can be generated in bulk semiconductor materials by various processes, such as light absorption, high electric field, carrier injection through a barrier, irradiation with high-energy particles. The non-equilibrium free carriers participate in the electrical transport. For instance, in the case of the excitation by light absorption, the supplementary current due to the non-equilibrium carriers is called photocurrent. An injection current is generated when carriers are injected over a Schottky barrier or a p–n junction. After the process which has generated the non-equilibrium carriers has ceased, the system returns to equilibrium due to the annihilation of the electron–hole pairs by recombination. The relaxation in time of the system toward thermodynamic equilibrium follows an exponential law [1]. If the carriers that recombine are both free (the electron in the conduction band and the hole in the valence band), their annihilation process is called band-to-band recombination. If one of the carriers is captured on a localized state (i.e., it has a fixed position in the semiconductor) and the other one is free, this is called recombination on localized states. By the recombination process, an amount of energy is released by the emission of either a photon (radiative recombination), or a phonon (non-radiative recombination), or a secondary electron (Auger recombination), etc.

Non-equilibrium carriers are strongly influenced by localized states (point defects, impurities, surface states, etc.). A neutral localized state can capture a non-equilibrium electron or hole and then the capture center becomes charged. In other words, a capture center is charged when it has a carrier localized on it. On the contrary, a dopant is neutral with the carrier localized on it, e.g., a phosphorus atom with $4 + 1$ valence electrons, or a boron atom with $4 - 1$ valence electrons, in a silicon crystal (four valence electrons per atom). The rate R_n of the capture process of electrons (with concentration n) on localized states (with concentration N_t) is given by the expression $R_\mathrm{n} = c_\mathrm{n} n N_\mathrm{t}$, where $c_\mathrm{n} = \varsigma_\mathrm{n} \tilde{v}_\mathrm{e}$ is the capture coefficient, ς_n is the capture cross-section, and $\tilde{v}_\mathrm{e} = \sqrt{3k_\mathrm{B}T/m_\mathrm{e}^*}$ is the thermal velocity of the electrons [1,2]. A similar expression can be written for the holes ($R_\mathrm{p} = c_\mathrm{p} p P_\mathrm{t}$). The capture process releases (thermal) energy to the lattice.

A carrier captured on a localized state can recombine with an opposite sign carrier, if this opposite sign carrier is subsequently captured on the same localized

state, or it can be released in the corresponding band (conduction band for electrons and valence band for holes). The capture of a second carrier with an opposite sign on the same localized state leads to the annihilation of the pair and it is called recombination on a localized state. The lifetime of a free non-equilibrium carrier, $\tau_n = 1/c_n N_t$ for electrons and $\tau_p = 1/c_p P_t$ for holes, is by definition the mean time during which the carrier is free before recombination.

If the captured carrier is released in the band, the capture center is called a trap. Then the capture process is called trapping and the release process is called detrapping. The detrapping rate is defined by analogy with the trapping rate, namely $R'_n = c_n n_t N_{ct}$ for electrons, where

$$N_{ct}(T) = N_c(T)\exp\left[-\frac{\Delta E_{tn}}{k_B T}\right] \equiv 2\left(\frac{m_e^* k_B T}{2\pi\hbar^2}\right)^{3/2}\exp\left[-\frac{E_c - E_{tn}}{k_B T}\right],$$

N_c being the effective density of states in the conduction band and ΔE_{tn} the depth of the trapping level into the band gap measured from the edge of the conduction band (the trap activation energy). A similar expression can be written for holes. If $R'_n > R_n$, the capture center acts as a trap and if $R_n > R'_n$, the capture center acts as a recombination center.

Figure 8.1 presents the transitions in bulk semiconductors. In case (a), the intrinsic absorption (transition 1) and the band-to-band recombination (transition 2) are illustrated. Case (b) shows the recombination processes on localized states as follows: on level L_{R1}, one electron is first captured and then it is annihilated by a hole capture, while on level L_{R2}, one hole is first captured and then it is annihilated by an electron capture. In case (c), the trapping and detrapping processes are presented for two kinds of traps, traps for electrons and traps for holes, respectively. The trapping of either an electron or a hole is

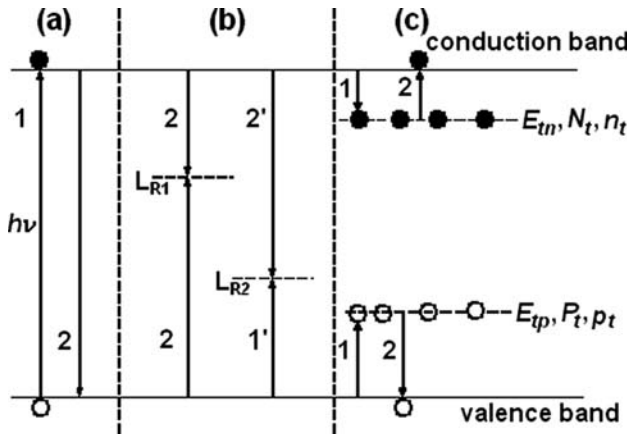


Fig. 8.1 Transitions in bulk semiconductors: (**a**) intrinsic absorption and band-to-band recombination; (**b**) recombination levels; (**c**) trapping levels

quoted 1 and the detrapping is quoted 2. A trapping level is characterized by the following parameters: the activation energy ($\Delta E_{tn} = E_c – E_{tn}$ for electron traps and $\Delta E_{tp} = E_{tp} – E_v$ for hole traps), the capture cross-section (ς_n for electrons and ς_p for holes), the trapping center concentration (N_t and P_t for electrons and holes, respectively), and the trapped electron (hole) concentration ($n_t \leq N_t$ and $p_t \leq P_t$). In other words, the trapped carrier concentration represents the concentration of the trapping centers occupied by electrons (holes).

The traps can be classified based on different criteria. Following the trapped charge, there are two kinds of traps: traps for electrons and traps for holes. Following the location, there are also two kinds of traps: bulk traps and surface/interface traps. The impurities, point defects, surface states, dangling bonds, and also localized stresses can act as trapping centers. The trapping levels are usually located in the band gap, but they can also exist in the conduction or in the valence band.

Let us discuss an example of each kind of trapping centers. A carbon atom that substitutes a silicon atom in a silicon crystal is an *impurity*. Both elements have the same valence and they crystallize in the same lattice. However, the carbon atom has a smaller effective radius and therefore it attracts electrons more strongly. This means that it can easily trap an electron. An arsenic vacancy in a gallium arsenide crystal is a *point defect*. This vacancy can trap electrons because the missing arsenic is pentavalent, while its nearest neighbors are trivalent gallium atoms. The nitrogen dioxide can be adsorbed at the silicon surface and is bonded there through Van der Waals forces. Then, it forms a *surface state* that can trap an electron (the nitrogen dioxide being an oxidant). Similar examples can be given for hole traps. A silicon atom located at the surface of a (111) silicon wafer is bounded to three other atoms, but the fourth orbital (oriented toward the exterior) is not bonded (*dangling bond*) and it can trap an electron to form a stable electronic configuration. A *local stress* can appear for instance at the (111) silicon–calcium fluoride interface. At room temperature (RT), the two lattices match very well. However, as their dilatation coefficients are different, local stresses will appear when the system is cooled down. Such a stress displaces the neighboring atoms and thus it creates a local potential that now acts as a trapping center.

In bulk semiconductors, trapping phenomena are dominated by the traps located in the volume of the crystals, such as point defects, impurities, and local stresses. At the same time, in nanocrystals these phenomena are dominated by the traps located at the surface/interface. The latter is due to the very large surface/volume ratio (of the order of 10^8 m^{-1}). Indeed, if we estimate the ratio of the number of atoms located at the surface toward the total number of atoms for a spherical nanodot (0D system), it is $N_S/N = 6a/d$, where a is the mean interatomic distance and d the nanodot diameter. For instance, a silicon nanodot ($a \approx 0.27$ nm) of about 3 nm diameter has practically half of the atoms located at the surface. For a cylindrical nanowire (1D system), the ratio becomes $N_S/N = 4a/d$ (again d is the nanowire diameter), while for a nanolayer (2D system), it becomes $N_S/N = 2a/d$ (d being the nanolayer thickness). Therefore, the number of atoms located at the surface/interface is of the same order of

magnitude as the total number of atoms. Any such atom can act like a trap or recombination center, so that the surface/interface of a nanosystem plays a major role in the non-equilibrium processes. Most of the surface/interface traps are adsorbed atoms/molecules, dangling bonds, or misfit-induced internal stresses [3,4].

In nanowires and nanolayers, one finds the same kinds of traps as in bulk semiconductors. In nanodots, a supplementary capture phenomenon appears on the quantum confinement (QC) levels [5]. The carriers captured this way are not localized at atomic scale, like the "classical" traps, but at nanodot scale (i.e., they can move inside the nanodot only). Thus, for "large" nanodots, with diameters much greater than 10 interatomic distances, this localization is weak. For "small" nanodots (hereafter called "quantum dots"), with diameters less than the order of magnitude of 10 interatomic distances, this localization is strong enough to play a significant role in the trapping–detrapping processes and/or the recombination on localized states. A specific characteristic of a quantum dot is that it has no real energy bands and no momentum conservation [6]. This is due to the small number of atoms that can be found in any direction.

For instance, if the nanodot diameter equals 10 interatomic distances, the maximum number of atoms that can be found in any direction is 11, i.e., 11 doubly degenerate states for each set of atomic quantum numbers (n, l, m), as it can be seen from Fig. 8.2. Such a set of distinct states does not form a proper band, due to the finite differences in energy (a proper band would need infinitely small differences). This means that the energy levels group themselves in sets (quasibands), separated by gaps much larger than the differences between the levels from the same quasiband. At the same time, the finite differences in momentum do not allow momentum conservation.

At the same time, the QC effects introduce supplementary levels, as the quantum dot surface acts like the wall of a quantum well. These levels are located over the last occupied level at absolute zero temperature (i.e., between

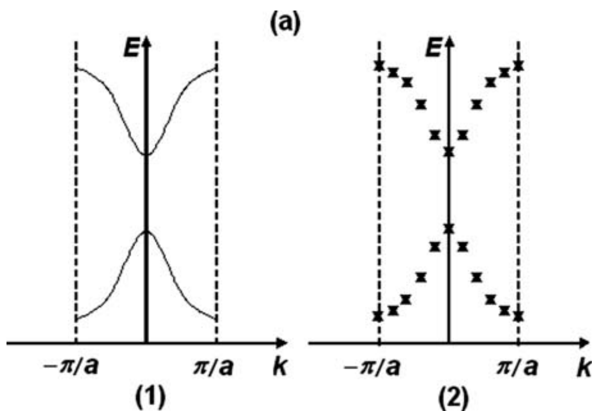


Fig. 8.2a Energy versus wavevector in (1) bulk semiconductor (continuous bands) and (2) quantum dot (discrete states, marked by x, separated by forbidden intervals)

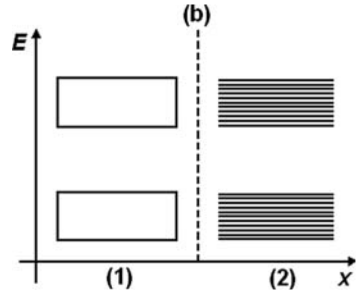


Fig. 8.2b Energy versus position in (1) bulk semiconductor (continuous bands) and (2) quantum dot (discrete states)

the quasi-valence and the quasi-conduction bands). There are two ways to make such levels act like traps. One way is to inject a carrier into a quantum dot by tunneling (there is no other way to make a carrier enter or leave a quantum dot, as its surface acts like a barrier). This supplementary carrier cannot recombine with an opposite sign carrier, as there are no real energy bands in a quantum dot to act like carrier reservoirs [7]. The carrier remains trapped until it leaves the quantum dot or recombines with an opposite sign carrier subsequently injected in the quantum dot. Another way to produce a non-equilibrium carrier is to excite an electron from a lower state into a higher one (producing at the same time a hole in the initial state). If the tunneling probabilities of the electron and hole are different, one of them leaves the quantum dot and then both will act like non-equilibrium carriers trapped in different dots [6].

The behavior of carriers in quantum dots also depends on their Coulomb interactions. The Coulomb repulsion between two equal charges inside a quantum dot becomes so important that it does not allow the simultaneous presence of more than one non-compensated charge in the quantum dot. This phenomenon is called "Coulomb blockade". Therefore, if several trapping centers for the same sign carriers are located in the same quantum dot, only one of them could be occupied. If traps of both signs are located in the same quantum dot, special complications could arise. Let us discuss the case of three trapping centers located in the same quantum dot, with the activation energies ΔE_{t1}, ΔE_{t2}, and ΔE_{t3} (by convention, $\Delta E_{t1} \leq \Delta E_{t2} \leq \Delta E_{t3}$). What happens if we try to charge all of them? If the first two centers (ΔE_{t1} and ΔE_{t2}) are traps for the same sign carriers, they cannot be simultaneously charged, whatever charge could be trapped on the third center. If the first two centers trap opposite sign carriers, all three centers can be charged simultaneously.

The discharge of the traps is also dependent on the charge signs. If the first and third centers trap charges with the same sign, the discharge of the three levels will happen in normal order, i.e., one by one, starting with the lowest energy. If the second and third centers trap charges with the same sign (opposite to the first center, see Fig. 8.3a), the discharge of the first level would imply the double charging of the quantum dot, which is forbidden by the Coulomb blockade (Fig. 8.3b). Consequently, the second level will be discharged

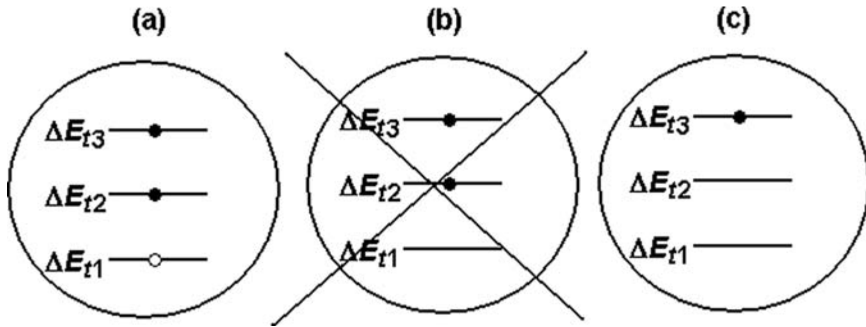


Fig. 8.3 The trapping–detrapping process in a quantum dot with three centers. (**a**) All traps are charged; (**b**) the lowest energy level cannot be discharged solely, due to the Coulomb blockade effect; (**c**) simultaneous discharge of the lowest two energy levels

simultaneously with the first one (Fig. 8.3c). This means that the experimental measurements will lead to the observation of a single experimental maximum corresponding to the first two levels, with the apparent activation energy $\Delta E_t \approx (\Delta E_{t1} + \Delta E_{t2})/2$ [8]. Meanwhile, the discharge of the third level will produce a "normal" maximum, with the real activation energy ΔE_{t3}.

The investigation of the trapping–detrapping phenomena in semiconductor materials and devices is essential for several processes. In the case of nanocrystalline semiconductors, specific effects appear. In the electrical transport through a quantum dot system, the current is reduced by the trapping of the carriers inside the quantum dots. In the case of the phototransport, the trapping of one type of carrier (electrons or holes) increases the lifetime of the opposite sign carriers and thus the recombination rate decreases. Therefore the photocurrent increases. Light absorption is always increased by the presence of the traps, while light emission depends on the radiative versus non-radiative contributions.

It is important to know both how to use the traps beneficially whenever possible and how to minimize the problems they can create. There are several specific applications of the traps. For example, in NROM (Nitride Read Only Memory) non-volatile memory cells, the traps in the silicon nitride (Si_3N_4) nanolayer are used for charge storage [9]. Also, the traps at the surface/interface of silicon nanodots are proposed for application in the memory devices [5,10,11]. On the other hand, traps can induce a reduction of the device reliability. For example, in the case of thin SiO_2 gates, the traps contribute to the wear out of the oxide through the weakening of the Si–O bond by the trapped electrons [12,13]. The study of p-channel MOSFETs proved that the application of a moderate temperature stress at negative bias induces a significant increase in interface trap concentration, which in turn produces an increase in the temperature instability of the transistor at a negative bias [14].

8.2 Classical Investigation Methods

Several methods with different applicability conditions have been used to study the trapping phenomena. They are generally related to the transient processes. The simplest type of transient process is represented by the *exponential decay* of a measured quantity Q, where the relaxation time τ is time-independent [1]:

$$Q(t) = Q_\infty + (Q_0 - Q_\infty)\exp(-t/\tau). \tag{8.1}$$

The quantity Q may be a photocurrent, a photoluminescence spectral intensity, a capacitance, etc. If the process described by Eq. (8.1) is due to the electron detrapping, the relaxation time τ becomes the carrier lifetime. For electrons, $\tau_n = 1/c_n N_{ct}$, where the electron capture coefficient is $c_n(T) = \varsigma_n(T)\,\tilde{v}_n(T) \equiv \varsigma_n(T)\cdot\sqrt{3k_B T/m_e^*}$, ς_n being the capture cross-section and $\tilde{v}_n$ being the thermal electron velocity, while

$$N_{ct}(T) = N_c(T)\exp\left[-\frac{\Delta E_{tn}}{k_B T}\right] \equiv 2\left(\frac{m_e^* k_B T}{2\pi\hbar^2}\right)^{3/2}\exp\left[-\frac{E_c - E_{tn}}{k_B T}\right], \tag{8.2}$$

where N_c is the effective density of states in the conduction band and ΔE_{tn} is the depth of the trapping level into the band gap (the trap activation energy). Similar expressions can be written for hole traps.

If there is a single zero-width trapping level and there is no recombination level (there is only band-to-band recombination), then by measuring the time decay of a quantity Q at different temperatures we can find the lifetime as function of temperature. If in addition ς_n is temperature-independent, then the plot of $\ln(\tau T^2)$ versus $1/T$ gives the trap activation energy ΔE_{tn}. Using ΔE_{tn}, we compute N_{ct} at a given temperature. Then, from the value of the lifetime τ_n and of the thermal electron velocity $\tilde{v}_n$ at the same temperature, we can also find the capture cross-section ς_n. At the same time, the signal amplitude $\Delta Q = Q_0 - Q_\infty$ is proportional to the trap concentration N_t. This is the simplest method to obtain information about traps from the experimental data. The analysis becomes more intricate if at least one of the following conditions is fulfilled: (i) the lifetime τ_n is time-dependent, (ii) we have strong retrapping, or (iii) there is more than a single zero-width trapping level.

A better investigation method is the *rate window* approach. When studying the decay process of the quantity Q at a constant temperature, one measures it at two consecutive moments, t_1 and t_2 ($t_1 < t_2$). This measurement is repeated at different temperatures. Then the difference $\Delta Q(T) \equiv Q(t_1, T) - Q(t_2, T)$ is represented as a function of temperature. As it can be seen from Eq. (8.1), at low temperatures the decay is slow, while at high temperatures it is fast, so that in both cases ΔQ is small. At intermediate temperatures, ΔQ will reach a maximum value (meaning $\partial\Delta Q/\partial T \equiv (\partial\Delta Q/\partial\tau)\cdot(\mathrm{d}\tau/\mathrm{d}T) = 0$). This maximum condition is related to

the trap parameters by the relation $\tau(T_m) = (t_1 - t_2)/\ln(t_1/t_2)$. The best choice for t_2 is a value large enough, when $Q(t_2) \approx Q_\infty$. Then, the moment at which the maximum value for ΔQ is obtained is simply $t_1 = \tau(T_m)$.

Due to the temperature dependence of the lifetime, another way to investigate the transient phenomena is through their *thermally stimulated behavior*. If one heats the sample at a constant rate $\beta = \mathrm{d}T/\mathrm{d}t$, the plot of the difference between the thermally stimulated quantity Q and its equilibrium value Q_∞ versus temperature gives information about the trap parameters. In order to facilitate the computation, the heating rate is chosen small enough to ensure a quasistatic process.

In the following, we will discuss some of the most commonly used methods. The *deep level transient spectroscopy* (DLTS) is a method, suitable for nonhomogeneous samples with a well-defined space charge region [15,16]. By measuring the transient capacitance of a junction, it is possible to determine the energy, concentration, and capture cross-section of the trapping centers from the junction. This is valid under the assumption that the trap concentration is much smaller than the dopant one. The capacitance transient is produced by the filling and thermal emptying of the traps.

The junction capacitance C is given by the formula

$$C = A\sqrt{\frac{\varepsilon_0\varepsilon_r e N_S}{2(U_D - U)}}, \tag{8.3}$$

where U is the applied bias, U_D the depletion bias, and N_S the total space charge density. The application of the DLTS method to investigate trapping phenomena is made as follows. Let us consider a p^+–n junction with n-type traps. The initial capacitance C_0 is measured at a given temperature under a constant reverse bias U_R that fixes the depletion layer width d_1 (see Fig. 8.4a). Then, the bias is abruptly cut off ($U = 0$ V). This narrows the depletion layer ($d_2 < d_1$, Fig. 8.4b), increases the junction capacitance, and traps majority carriers from the n region. If we reapply the reverse bias U_R, the depletion width is increased with respect to the initial value ($d_3 > d_1$, Fig. 8.4c), due to the trapped carriers,

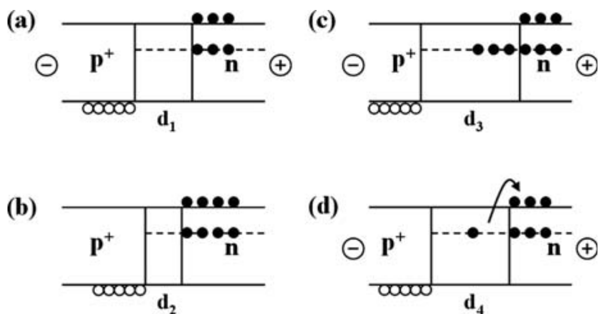


Fig. 8.4 The time evolution of the space charge region during DLTS measurements: (**a**) initial state (reverse bias U_R); (**b**) majority carrier pulse (zero bias); (**c**) start of the transient (reverse bias U_R); (**d**) space charge region during thermal detrapping

and therefore the capacitance is decreased with an amount $\Delta C_0 < 0$ with respect to its initial value C_0. The thermal discharge of the traps reduces the trapped charge and the depletion layer width ($d_1 < d_4 < d_3$, Fig. 8.4d) so that the capacitance transient has the form $\Delta C = C(t) - C_0 \equiv \Delta C_0 \exp(-t/\tau)$.

The measurements are repeated at different temperatures. Using the rate window method, the lifetime is determined from the maximum value of the transient capacitance. Using the expression of the lifetime (see the exponential decay method), the trap parameters can be determined. In a similar way, minority carrier trap parameters can be determined. In this case, $\Delta C_0 > 0$. We can also use n^+–p junctions; then the roles of the electrons and holes are reversed.

The method can be applied for high trap concentrations if one replaces the measurement of transient capacitance under constant voltage with the measurement of transient voltage under constant capacitance (*constant capacitance voltage transient* – CCVT) [17]. These methods are widely used to investigate trapping phenomena in semiconductor materials, where good p^+–n or n^+–p junctions can be fabricated. However, neither of these methods is suitable for nanocrystalline systems. Indeed, either such systems do not have a space charge region (the case of 0D systems) or the transient capacitance is too small for the measurement possibilities (1D and 2D systems).

In the *photoinduced current transient spectroscopy* (PICTS), the charging of the traps is made by illuminating the sample. The rate window method is applied to the photocurrent, the time t_2 being chosen such as to have the photocurrent at that moment practically equal with the dark current, i.e., $\Delta j(t_1) = j_{\text{photo}}(t_1) - j_{\text{dark}}$ [18,19,20,21]. If one considers a (neutral) defect level at thermal equilibrium, the absorption of a photon by the defect will imply the emission of an electron to the conduction band and the emission of a hole to the valence band. The emission rates are equal:

$$n_t c_n N_c \exp\left(-\frac{E_c - E_t}{k_B T}\right) = (N_t - n_t) c_p P_v \exp\left(-\frac{E_t - E_v}{k_B T}\right). \tag{8.4}$$

Using the previous definition for the electron and hole lifetimes, it results that, under illumination,

$$\frac{n_t}{\tau_n} - (N_t - n_t) c_n \Delta n = \frac{N_t - n_t}{\tau_p} - n_t c_p \Delta p, \tag{8.5}$$

where Δn and Δp are the non-equilibrium photogenerated carrier concentrations. The current density generated by the emptying of the level is

$$j(t) = \Gamma\left(\frac{n_t(t)}{\tau_n} + \frac{N_t - n_t(t)}{\tau_p}\right), \tag{8.6}$$

where Γ is a geometrical constant dependent on the light penetration depth. By using relations (4, 5, 6), one obtains

$$\Delta j(t) = j(t) - j(\infty) = \Gamma\left(\frac{1}{\tau_n} - \frac{1}{\tau_p}\right) \times \left[\frac{1}{1 + (\tau_p/\tau_n)(1 + \tau_n c_p \Delta p)/(1 + \tau_p c_n \Delta n)} - \frac{1}{1 + (\tau_p/\tau_n)}\right] \times N_t \exp\left(-\frac{t}{\tau}\right), \quad \frac{1}{\tau} = \frac{1}{\tau_n} + \frac{1}{\tau_p}. \tag{8.7}$$

For high light intensities ($\Delta n \approx \Delta p$; $\tau_n c_n, \tau_p c_p >> 1/\Delta n$), $\Delta j(t)$ becomes

$$\Delta j(t) = \Gamma\left(\frac{1}{\tau_n} - \frac{1}{\tau_p}\right)\left[\frac{1}{1 + (c_p/c_n)} - \frac{1}{1 + (\tau_p/\tau_n)}\right] N_t \exp\left(-\frac{t}{\tau}\right). \tag{8.8}$$

If the defect is simply an electron trap ($\tau_p = \infty$, $c_p = 0$), Eq. (8.8) takes the same form as Eq. (8.1), with $\Delta j_0 = \Gamma N_t/\tau_n$.

This method is successfully applied to high resistivity materials. Because the factor Γ includes geometrical information, as well as information about the trap concentration, the latter cannot be easily determined from the measurements. The method can be applied to nanocrystals, but here too it raises difficulties to the determination of the trap concentration.

The *thermally stimulated currents* (TSC) method is mainly used for high resistivity semiconductors and therefore it is also appropriate to the investigation of the trapping phenomena in nanocrystalline semiconductors [20,22]. In this method, the first step is to fill the traps at low-enough temperature T_0, in order to reduce the detrapping process as much as possible. The traps are filled by illuminating the sample with (monochromatic) light in the absorption band. When we study thin films, it is convenient for the modeling to have the penetration depth L_λ greater than the film thickness d. The photogenerated carriers diffuse with different velocities ($v_n \neq v_p$) into the sample. It is also convenient to have the bipolar diffusion length L_D greater than d (in the case of thin films). If both L_λ and L_D are greater than d, and if the illumination time is chosen sufficiently long and the light intensity sufficiently high, the traps from the thin film will be uniformly filled. If not, their filling will decrease with the depth z (measured from the illuminated surface). After switching off the light, a constant bias is applied and the sample is heated up at a constant rate $\beta = dT/dt$, small enough to ensure a quasistatic process.

During the heating, a current is measured as a function of temperature. This current is due to the contribution of the (non-equilibrium) detrapped carriers and of the thermally excited equilibrium ones, both moving under the externally applied field. In order to separate the two contributions, another measurement is made under the same thermal and electrical conditions, but in the dark (without filling the traps by illumination). Then, the thermally stimulated current

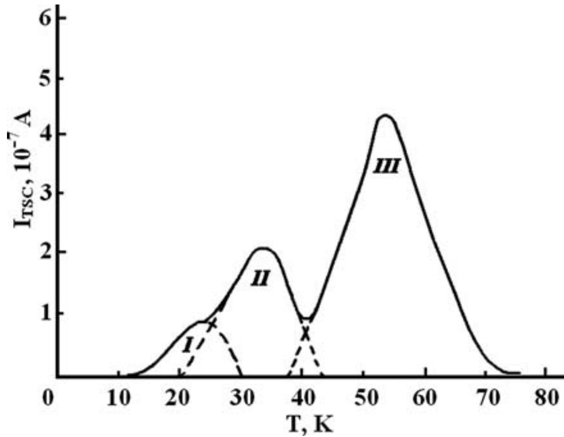


Fig. 8.5. Typical TSC discharge curve, measured on porous silicon [20]. Reprinted from Solid State Electronics, **46** (1), O. V. Brodovoy, V. A. Skryshevsky, and V. A. Brodovoy, "Recombination properties of electronic states in porous silicon", 83–87, Copyright © 2002, with permission from Elsevier

ΔI_{TSC} is defined as the difference between the current after the photoexcitation and the dark current. The thermally stimulated current versus temperature curve presents maxima and/or shoulders, corresponding to the different trapping levels (see Fig. 8.5, Ref. [20]). The curve maxima are related to the trap energies.

Let us consider first a curve with a single maximum. If one takes into account the increasing part of the curve, this part is described by an exponential law, $\Delta I_{TSC} \propto \exp(-\Delta E_t/k_B T)$, called Arrhenius law, so that the trap activation energy can be determined from $\ln(\Delta I_{TSC})$ as function of $1/T$. The area under the maximum of the ΔI_{TSC} versus T curve is proportional to the trap concentration (and the applied bias). When the curve contains several well-separated maxima, this procedure can be applied separately for each maximum. However, if two (or more) maxima are too close to each other, this method does not work anymore. In that case, to separate the contributions of different trapping levels, a fractional heating procedure has to be performed. For this, after the same cooling and illumination conditions, the heating is stopped at the first current maximum or shoulder. The temperature is kept constant until the current reaches the dark current value (the corresponding trapping level is discharged); then the sample is cooled down again (without any more illumination) and heated up to the next maximum or shoulder and so on. In this way, separate Arrhenius curves are obtained for each trapping level, and their activation energies can be obtained from these curves (Fig. 8.6, Ref. [23]). To obtain the trap concentrations as well, a more refined modeling of the process is needed.

The *thermally stimulated depolarization currents* (TSDC) method is also used for high-resistivity semiconductors [23,24,25]. Here the cooling is made under a constant bias, and the heating starts after switching off the bias (no optical excitation). The interpretation of the $I_{TSDC}(T)$ curve is similar to the TSC one, as the curves are similar in shape (see Fig. 8.6). In the TSDC spectrum, the highest temperature maximum is due to the capacitor-like behavior of the sample so that it has the same activation energy as the (dark) conductivity.

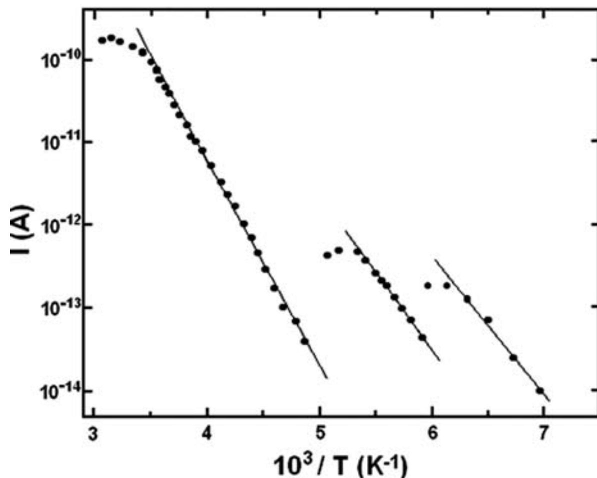


Fig. 8.6 Typical fractional heating TSDC curve for porous silicon [23]. Reprinted from Thin Solid Films **325** (1–2), M. L. Ciurea, I. Baltog, M. Lazar, V. Iancu, S. Lazanu, and E. Pentia, "Electrical behaviour of fresh and stored porous silicon films", 271–277, Copyright © 1998, with permission from Elsevier

The *optical charging spectroscopy* (OCS) is a zero bias method [26,27,28]. The first step is the same as for TSC, i.e., the filling of the traps is made by illuminating the sample at low temperature, using (monochromatic) light in the absorption band. The photogenerated carriers diffuse into the sample with different velocities and some of them are trapped, while the others recombine. The trapped carriers generate a frozen-in electric field, linearly dependent on the trap concentrations. Unlike the TSC method, here the heating is made without applying any external bias. During heating, the detrapped carriers move under the field of the still trapped ones, generating a discharge current. Therefore, the dependence of the current on the trap concentrations is stronger than that in the other methods.

As the OCS method is less known, its modeling [8,28] is shortly presented hereafter. The model supposes that (a) the sample has sandwich configuration (semitransparent top electrode/nanocrystalline film/crystalline semiconductor/bottom Ohmic electrode) and practically all the traps are located in the nanocrystalline film; (b) the heating is quasistatic ($\beta = \mathrm{d}T/\mathrm{d}t$ is constant and small enough to ensure this regime); and (c) only zero-width trapping levels are considered. When heating at a constant rate, the time and temperature dependences of the involved quantities are related by the condition $\partial/\partial t = \beta \cdot \partial/\partial T$. Then, the temperature dependence of the trapped carrier concentrations during the heating is given by the following equations:

$$\frac{\partial n_{\mathrm{t}i}(z,T)}{\partial T} = \frac{1}{\beta} c_{\mathrm{n}i}(T)\{[N_{\mathrm{t}i}(z) - n_{\mathrm{t}i}(z,T)]\Delta n(z,T) - N_{\mathrm{ct}i}(T) n_{\mathrm{t}i}(z,T)\}, \quad (8.9)$$

$$\frac{\partial p_{\mathrm{t}k}(z,T)}{\partial T} = \frac{1}{\beta} c_{\mathrm{p}k}(T)\{[P_{\mathrm{t}k}(z) - p_{\mathrm{t}k}(z,T)]\Delta p(z,T) - P_{\mathrm{vt}k}(T) p_{\mathrm{t}k}(z,T)\}. \quad (8.9')$$

The right hand side of Eqs. (8.9) and (8.9′) represents the difference between the trapping and detrapping rates. The non-equilibrium carrier concentrations detrapped from the levels i (for electrons) and k (for holes) result from the equations describing the total free carrier concentrations:

$$\begin{aligned}\frac{\partial}{\partial T}n(z,T) &\equiv \frac{\partial}{\partial T}\left[n_0(T)+\sum_i \Delta n_i(z,T)\right]\\ &= \frac{1}{\beta}\sum_i\left[c_{\mathrm{n}i}(T)N_{\mathrm{ct}i}(T)n_{\mathrm{t}i}(z,T)-\frac{\Delta n_i(z,T)}{\tau_{\mathrm{n}i}(T)}\right]\end{aligned} \tag{8.10}$$

$$\begin{aligned}\frac{\partial}{\partial T}p(z,T) &\equiv \frac{\partial}{\partial T}\left[p_0(T)+\sum_k \Delta p_k(z,T)\right]\\ &= \frac{1}{\beta}\sum_k\left[c_{\mathrm{p}k}(T)P_{\mathrm{vt}k}(T)p_{\mathrm{t}k}(z,T)-\frac{\Delta p_k(z,T)}{\tau_{\mathrm{p}k}(T)}\right],\end{aligned} \tag{8.10′}$$

n_0 and p_0 being the equilibrium carrier concentrations and Δn_i, Δp_k the concentrations of the carriers detrapped from levels i and k, respectively. If the heating regime is quasistatic, $\partial n/\partial T$ and $\partial p/\partial T$ vanish, so that the non-equilibrium (detrapped) carrier concentrations are

$$\Delta n(z,T) \equiv \sum_i \Delta n_i(z,T) = \sum_i \tau_{\mathrm{n}i}(T)c_{\mathrm{n}i}(T)N_{\mathrm{ct}i}(T)n_{\mathrm{t}i}(z,T) \tag{8.11}$$

$$\Delta p(z,T) \equiv \sum_k \Delta p_k(z,T) = \sum_k \tau_{\mathrm{p}k}(T)c_{\mathrm{p}k}(T)P_{\mathrm{vt}k}(T)p_{\mathrm{t}k}(z,T), \tag{8.11′}$$

and Eqs. (8.9) and (8.9′) become

$$\begin{aligned}&\left\{1+\sum_{i'}\tau_{\mathrm{n}i'}c_{\mathrm{n}i'}(T)\left[N_{\mathrm{t}i'}(z)-n_{\mathrm{t}i'}(z,T)\right]\right\}\frac{\partial}{\partial T}n_{\mathrm{t}i}(z,T)\\ &\quad= -\frac{1}{\beta}c_{\mathrm{n}i}(T)N_{\mathrm{ct}i}(T)n_{\mathrm{t}i}(z,T),\end{aligned} \tag{8.12}$$

$$\begin{aligned}&\left\{1+\sum_{k'}\tau_{\mathrm{p}k'}c_{\mathrm{p}k'}(T)\left[P_{\mathrm{t}k'}(z)-p_{\mathrm{t}k'}(z,T)\right]\right\}\frac{\partial}{\partial T}p_{\mathrm{t}k}(z,T)\\ &\quad= -\frac{1}{\beta}c_{\mathrm{p}k}(T)P_{\mathrm{vt}k}(T)p_{\mathrm{t}k}(z,T).\end{aligned} \tag{8.12′}$$

These equations can be solved only numerically, except for the case of the weak retrapping, $\tau_{\mathrm{t}}c_{\mathrm{t}}N_{\mathrm{t}} << 1$, when analytical solutions can be found [28].

Using the solutions of Eqs. (8.9) and (8.9′), the frozen-in field can be computed by the expression

$$E_z(z,T) = \frac{e}{\varepsilon_0 \varepsilon_\mathrm{r}} \int_0^z \left[\sum_k p_{\mathrm{t}k}(z',T) - \sum_i n_{\mathrm{t}i}(z',T) \right] \mathrm{d}z'. \tag{8.13}$$

Five different contributions to the discharge current can be taken into account: the non-equilibrium and the equilibrium carrier conduction currents, the displacement and the tunneling currents, and the diffusion one. It can be seen from Eq. (8.13) that the frozen-in electric field depends linearly on the trap concentrations, so that the current dependence on the trap concentrations is quadratic for the non-equilibrium carrier conduction current and exponential for the tunneling current (both increasing the sensitivity of the OCS method with respect to the previously discussed methods). From the fitting of the experimental data, several trap parameters can be obtained: trap activation energies, trap concentrations, capture cross-sections, and non-equilibrium carrier lifetimes.

The methods presented in this paragraph can be easily applied to the case of bulk semiconductors. From the previous discussions, it results that some of them raise difficulties when applied to the nanocrystalline semiconductors (and devices). Therefore, the trapping phenomena can also be investigated by means of different non-conventional methods, with a partial determination of the trap parameters. In the following, several conventional and non-conventional methods applied to the most investigated nanocrystalline semiconductors will be presented.

8.3 Applications: Classical and Non-conventional Methods

Modern electronics studies more and more nanocrystalline materials and structures, where the trapping phenomena play a very important role. Presently, more than 80% of microelectronic devices are fabricated from silicon-based structures. Therefore, Si is the most investigated semiconductor for nanoscale applications.

The traps in nanocrystalline silicon (nc-Si) were proposed as tools for the quantum computers and memory devices [5]. For instance, a floating gate memory device was made from a SiO_2/nc-Si dots/SiO_2 tunneling layer/Si structure, presented in Fig. 8.7. This structure is fabricated on an (100) p-type Si wafer, with 8–10 Ω cm resistivity. The upper oxide has 41 nm thickness, while the tunneling one has 2 nm. Three kinds of samples, A, B, and C, were prepared. Sample A, for the memory device, had nc-Si dots with 7 ± 1 nm diameter. The nanodot density was 1.1×10^{11} cm^{-2}. The nanodots were coated with silicon nitride (SiN_x) film having a thickness of 1 nm. Sample B had uncoated nc-Si dots (with a diameter of

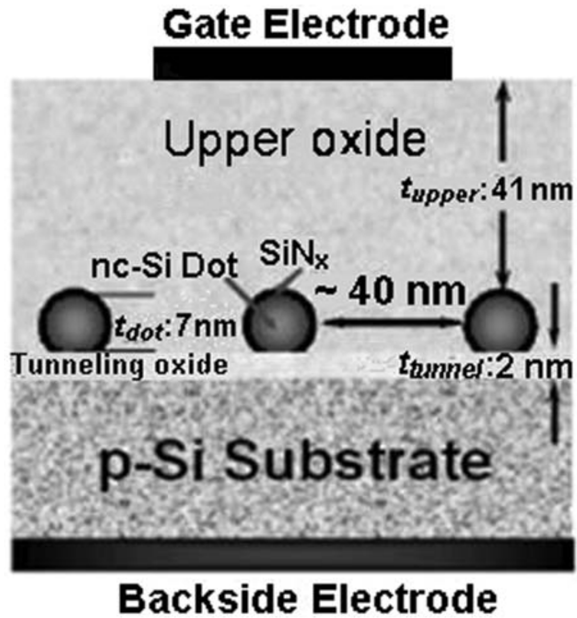


Fig. 8.7 The structure of the nitrided nc-Si dot memory [5]. Reused with permission from Shaoyun Huang and Shunri Oda, Applied Physics Letters, **87**, 173107, 2005. Copyright © 2005, American Institute of Physics

8 ± 1 nm and the same density), while sample C was prepared without nc-Si dots. The different samples were investigated for comparison.

The displacement current versus gate voltage characteristics measured on each sample under identical gate-voltage scan rate at RT are presented in Fig. 8.8. The forward current for sample A presents a well-marked maximum, while the reverse current has practically no maximum at all. Both forward and reverse currents for sample B exhibit marked maxima. For sample C (without nanodots), both currents are practically null. This behavior is explained by the band structure shown in Fig. 8.9. In the forward current regime (Fig. 8.9a), the electrons tunnel through the ultrathin SiO_2 layer (1) into the nc-Si dots (2). Once there, they are trapped on the QC levels (I), ensuring a fast charging (writing process). Part of these electrons trapped on the QC levels will then be captured onto the traps (II) located at the nc-Si/SiN_x interface (3). The reverse current regime represents the erasure process (Fig. 8.9b). During this process, the electrons trapped on the QC levels (I) are easily freed back to the substrate, while those trapped on the interface traps (II) need a larger erasure bias (or else a larger erasure time) to overcome the barrier and thus to be detrapped. The carriers captured on the QC levels are not localized at the atomic scale like the "classical" traps, but at the nanodot scale, i.e., they can freely move inside the nanodot (as was stated in Introduction). This work represents an original method to use two trapping processes on the same nanodot, on QC levels and on "classical" interfacial traps, to achieve a good performance of the device.

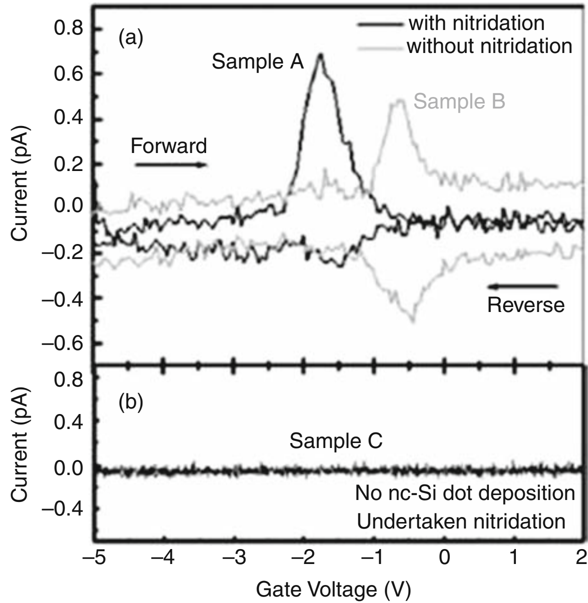


Fig. 8.8 Displacement current characteristics. (**a**) Charge/discharge current peaks of samples A and B; (**b**) no current peaks in sample C [5]. Reused with permission from Shaoyun Huang and Shunri Oda, Applied Physics Letters, **87**, 173107, 2005. Copyright © 2005, American Institute of Physics

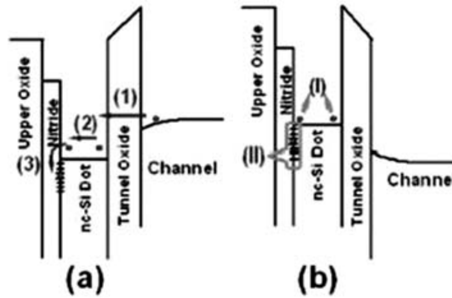


Fig. 8.9 (**a**) Writing and (**b**) erasure processes in nitrided nc-Si dot-based memory devices. (1) Direct tunneling from channel to nc-Si dot; (2) trapping on QC levels; (3) trapping at nc-Si/SiN_x interface; (I) QC levels; (II) interface traps [5]. Reused with permission from Shaoyun Huang and Shunri Oda, Applied Physics Letters, **87**, 173107, 2005. Copyright © 2005, American Institute of Physics

Another remarkable application of the carrier trapping on the QC levels is the fabrication of a light source from a floating gate MOSFET [29]. An array of quantum dots (2–4 nm diameter) is inserted in the oxide layer, very close to the substrate. By applying a positive bias on the gate, electrons are injected in the quantum dots (by Fowler–Nordheim tunneling) and are trapped on the QC levels (Fig. 8.10a). When the sign of the gate bias is reversed, holes are injected into the quantum dots (by Coulomb field-enhanced Fowler–Nordheim tunneling) and they either recombine with the trapped electrons or form excitons

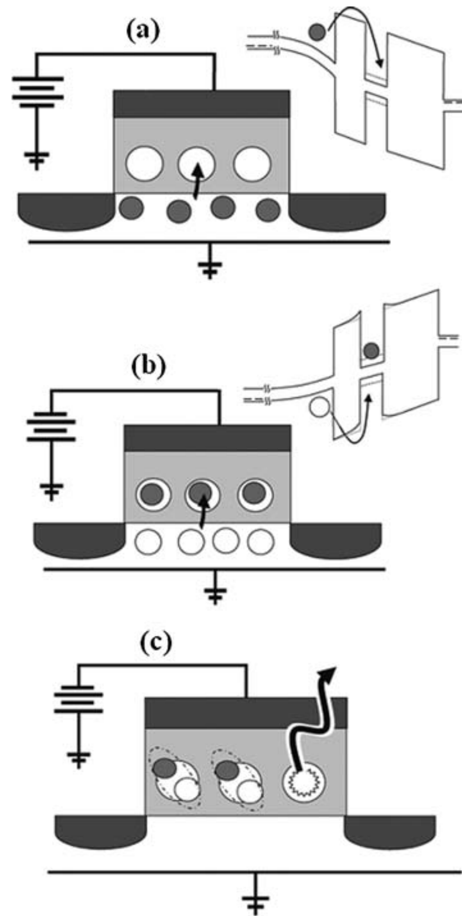


Fig. 8.10 (**a**) Electron injection in the quantum dots; (**b**) hole injection in the quantum dots; (**c**) exciton formation and radiative recombination process [29]. Reprinted by permission from Macmillan Publishers Ltd: Nature Materials, R. J. Walters, G. I. Bourianoff, H. A. Atwater, "Field-effect electroluminescence in silicon nanocrystals" **4**, 143, 2005. Copyright © 2005

(Fig. 8.10b). The radiative recombination of the excitons (Fig. 8.10c) gives the electroluminescent signal. Due to the radiative lifetime of the excitons (about 100 μs), the maximum efficiency of the electroluminescence is attained for a gate bias frequency of about 10 kHz.

The importance of the carrier trapping on the QC levels of a nanodot, or else at the nanodot interface, was also demonstrated by AFM measurements [30], which represent a non-conventional method to investigate the trapping phenomena. For this, two kinds of samples were studied by comparison. Both samples were prepared by ion implantation in wet thermally grown SiO_2 films, followed by annealing at 1100°C in vacuum. The first sample is made by Si^+ ion implantation. By annealing, the implanted silicon forms nanocrystals (2–6 nm diameter), as it can be seen from the AFM image in Fig. 8.11a.

The second sample uses implantation by Ar^+ ions followed by a similar annealing process, to produce similar implantation defects. No nanocrystals

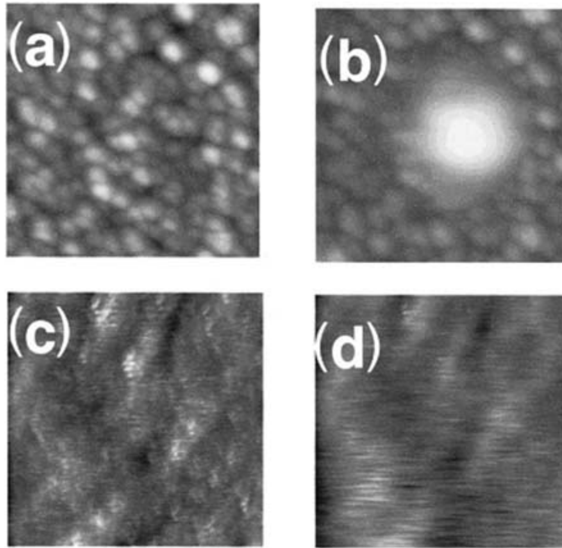


Fig. 8.11 AFM image of a SiO_2 film containing Si nanocrystals made by Si^+ ion implantation and annealing: (**a**) before charging and (**b**) after charge transfer [30]. Lateral size of the images is 5 μm: black to white (vertical) scale is 15 nm for (**a**) and 25 nm for (**b**). Similar images for Ar^+ ions' implantation: (**c**) before charging and (**d**) after charge transfer [30]. Lateral size for both images is 1 μm: black to white (vertical) scale is 1.5 nm. Reused with permission from E. A. Boer, M. L. Brongersma, H. A. Atwater, R. C. Flagan, and L. D. Bell, Applied Physics Letters, **79**, 791, 2001. Copyright © 2001, American Institute of Physics

appear in this case (Fig. 8.11c). The AFM tip is then used to charge both samples. The charge localization is checked by subsequent AFM imaging (Figs 8.11b, d). The image (b) shows a bright region of localized charge in sample A that appears like a surface prominence. The brightness decreases in time and practically disappears after about 600 s. No such region appears in sample B, where one finds similar defects, but no nanocrystals. Therefore, one can conclude that the trapped charge is localized in nanocrystals or at their interface with the SiO_2 matrix.

Another non-conventional method to investigate the traps is the study of the Coulomb blockade spectrum of a nanodot prepared in a silicon nanowire [31]. A silicon nanowire (20×30×200 nm) was made by the etching of a silicon-on-insulator (SOI) film with a weak As doping (10^{18} cm^{-3}). A local thermal oxidation was then performed, followed by the deposition of a supplementary oxide layer. Two samples were prepared: one with 2 nm thermal oxide and 8 nm deposited oxide (10 nm total thickness), the other with 4 nm thermal oxide and 20 nm deposited oxide (24 nm total thickness). Over the oxide layer, a polysilicon gate was deposited, together with 50 nm wide Si_3N_4 spacers on each side of the gate. Then, a second doping of the nanowire was made (by implantation), increasing the As concentration to 4×10^{19} cm^{-3}. This way the region under the gate remains

weakly doped and behaves like a nanodot (due to the lateral potential barriers that appear). To evidence the Coulomb blockade effects, a third sample was fabricated from highly doped SOI (10^{19} cm^{-3}). Only thermal oxide (4 nm thickness) was grown under the gate. Because no supplementary doping was performed, no spacers were needed in this case. This system acts like a classical MOSFET, while the first two samples form single electron transistors (SET).

The dependence of the drain–source conductance on the gate voltage is presented in Fig. 8.12 for all the three samples. The smooth (classical) characteristics taken at RT are replaced by the oscillations produced by the Coulomb blockade effect at low temperature (under 20 K, when the Coulomb blockade energy is larger than the thermal agitation energy). The period of these oscillations is related to the overlap between the nanowire and the gate. The Coulomb blockade oscillations in the thin gate oxide sample (4 nm) are randomly distributed. This is due to the fact that the Coulomb blockade in a nanowire without nanodot is induced by the gate potential only. On the

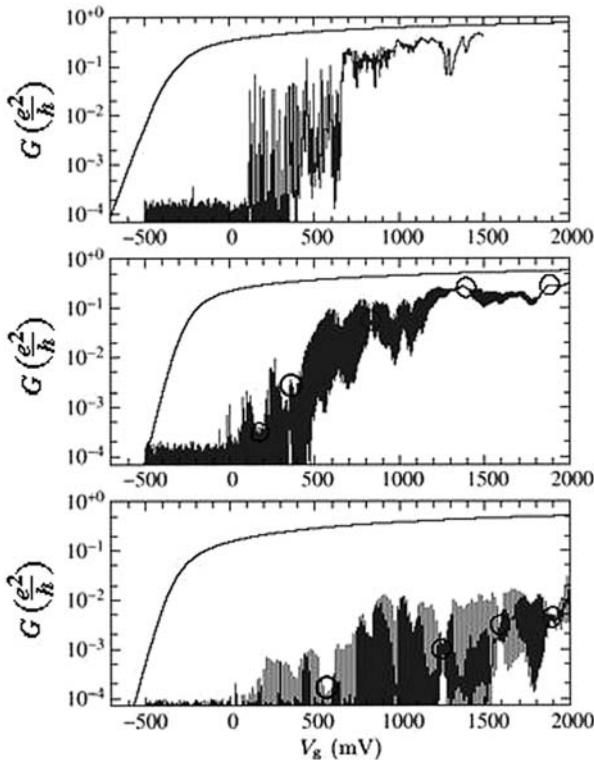


Fig. 8.12 Drain-source conductance dependence on the gate voltage for three different gate oxide thickness, from up to down: 4, 10, and 24 nm. The smooth curves are taken at RT, while the rapidly oscillating ones at 20 K [31]. Reused with permission from M. Hofheinz, X. Jehl, M. Sanquer, G. Molas, M. Vinet, and S. Deleonibus, “Individual charge traps in silicon nanowires”, Eur. Phys. J. B **54**, 299–307 (2006)

contrary, the oscillations in the other two samples are regular, due to Coulomb blockade in a well-defined nanodot (as prepared). A few anomalous decreases of the oscillation amplitude, marked with circles, appear in these curves. These anomalies can be attributed to the electrostatic interaction between the charges located in the nanodot and the charges trapped by dopant sites in the oxide or the nanowire. An intricate numerical analysis based on this assumption proved that these traps are located near or inside the wire.

Up to now we have discussed traps in silicon nanodots (0D systems). Let us analyze silicon nanowires (1D systems) as well. The traps from nanocrystalline porous silicon (nc-PS) were studied by means of the OCS method [27,28]. The nc-PS films are formed by a nanowire network, with diameters of 2–4 nm [23,32]. The excitation was made with strongly absorbed light, with wavelength $\lambda = 0.5\,\mu\text{m}$.

The results of the OCS measurements (dotted line), together with the theoretical curve obtained by modeling using Eqs. (8.11, 8.12, 8.13) (solid line), are presented in Fig. 8.13 [28]. The modeling allowed the determination of the trap parameters (trap activation energies and concentrations, capture cross-sections, and detrapped carrier lifetimes). These values are not presented here because this tutorial focuses on the methods and their applications, not on numerical values of the parameters.

As the maxima and shoulders are not well enough separated to determine the trap activation energies from their increasing parts, the fractional heating procedure was used. The maxima Nos. 1, 3, and 4 are "normal" and allowed the determination of the trap activation energies. However, the broad shoulder No. 2 could not be experimentally resolved (not even by fractional heating) so that the modeling was necessary. The theoretical curve obtained by modeling resolved the shoulder No. 2 into the maxima Nos. 2′ and 2″. The trap activation energies are proportional with the temperature of the corresponding maxima ($\Delta E_{t1} < \Delta E_{t2'} < \Delta E_{t2''} < \Delta E_{t3} < \Delta E_{t4}$). The maximum quoted F is an exception to this rule. Its apparent activation energy is much higher than ΔE_{t4}. This exception can be explained as follows. From Fig. 8.13, one can see that maxima Nos. 3 and 4 have opposite signs, meaning that they correspond to

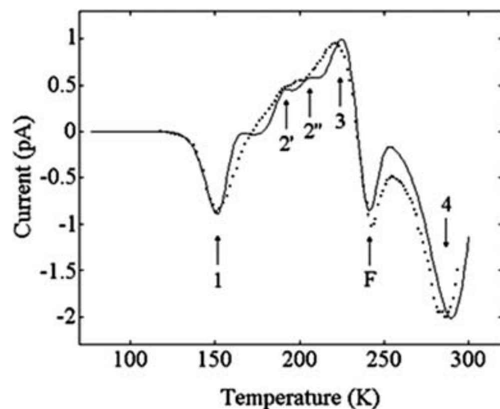


Fig. 8.13 Fitting of the OCS results for fresh nc-PS sample: *solid line* – model; *dotted line* – experimental data [28]. Reused with permission from Vladimir Iancu, Magdalena Lidia Ciurea, and Mihai Draghici, Journal of Applied Physics, **94**, 216, 2003. Copyright © 2003, American Institute of Physics

traps for opposite sign carriers. Therefore, during the discharge of the trapping level No. 3, the frozen-in electric field changes sign. As in nc-PS, the main discharge current is a conduction one; it will change sign together with the field. Because this change of sign happens before the level No. 3 is completely empty, the end of its discharge appears like a false maximum (F).

Similar measurements were made on oxidized samples, for the same wavelength [33]. The maxima/shoulders Nos. 1, 2 (i.e., 2′ and 2″), and 3 are flattened, proving that they should correspond to surface trapping centers, while the maximum No. 4 corresponds to volume centers.

The TSDC measurements [23] give the same trap activation energies as the first three maxima obtained from OCS (see Fig. 8.6). Maximum No. 2 cannot be resolved and maximum No. 4 does not appear, proving that this method is not sensitive enough.

As a first example of 2D systems, we investigated multi-quantum well (MQW) structures formed by a set of 50 bilayers of nc-Si and CaF_2, 1.6 nm thickness each, deposited on a silicon substrate, $(nc\text{-}Si/CaF_2)_{50}$ [34]. A surprising behavior appears in the OCS measurements [8,35]. The zero curve, taken in dark (without illumination at low temperature), presents two spikes (Fig. 8.14, curve *a*). The first spike also appear in the OCS curve (Fig. 8.14, curve *b*), while the second one is reduced to a shoulder. A supplementary maximum appears close to the second spike. All of them were well evidenced in fractional heating measurements.

Because the spikes appear in the zero curve, we infer that they are due to the misfit stresses that appear at the nc-Si/CaF_2 interfaces during the cooling. These

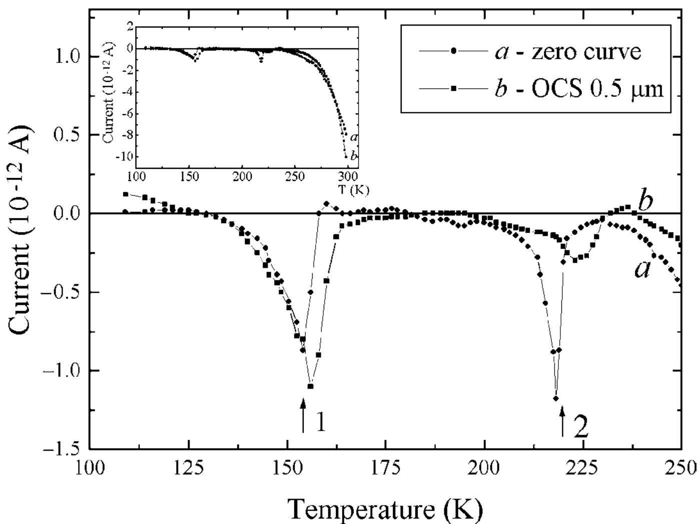


Fig. 8.14 OCS discharge current in MQW structure: (*a*) zero (no illumination) curve; (*b*) OCS ($\lambda = 0.5\,\mu m$) curve [8]. Reprinted from Solid State Electronics, M. L. Ciurea, V. Iancu, and R. M. Mitroi, "Trapping Phenomena in Silicon-Based Nanocrystalline Semiconductors", **51**, 1328–1337, 2007, Copyright © 2007, with permission from Elsevier

Table 8.1 Parameter values for the MQW trapping levels [8]

Maximum number	Maximum type	ς (10^{-18}cm^2)	N_t (P_t)(10^{14} cm^{-3})	τ(ns)	ΔE_t (eV)	
					Model	Exp.
1	nS	1.70	66.00	400	0.30	0.30
2	nS	0.41	26.00	400	0.42	0.42
3	pS	1.00	0.29	180	0.44	0.44
4	nS	1.50	55.00	400	0.72	0.75

Reprinted from Solid State Electronics, M. L. Ciurea, V. Iancu, and R. M. Mitroi, "Trapping Phenomena in Silicon-Based Nanocrystalline Semiconductors", **51**, 1328–1337, 2007, Copyright © 2007, with permission from Elsevier.

stresses act as traps and their filling depends on the cooling rate. The theoretical fit of the OCS curve, also made by using Eqs. (8.11, 8.12, 8.13), is shown in Fig. 8.15 [8]. To give a numerical example, we also list in Table 8.1 the trap parameters that result from the fit.

Another example of a 2D system is represented by the channel of a CMOS transistor with high ϵ_r gate dielectric material (also called high κ dielectric). The traps in high κ dielectrics can introduce a significant reduction of the apparent electron mobility through the charge accumulation in the dielectric [36]. The apparent mobility, determined from the pulse I_d–V_g measurements (in the μs or ns range), can be up to 27% smaller than the real one. Indeed, if a charge q_t is trapped in the dielectric, the transistor threshold voltage V_T is increased with $\Delta V_T = q_t/C_t$, where C_t is the trapped charge capacitance with respect to the substrate. Then, the apparent mobility is $\mu_a = \mu(1 - \partial\Delta V_T/\partial V_g)$ (μ is the real mobility and V_g is the gate voltage). Consequently, only fast trapping processes can produce significant effects in high-frequency transistors, otherwise the trapped charge cannot follow the gate voltage variations.

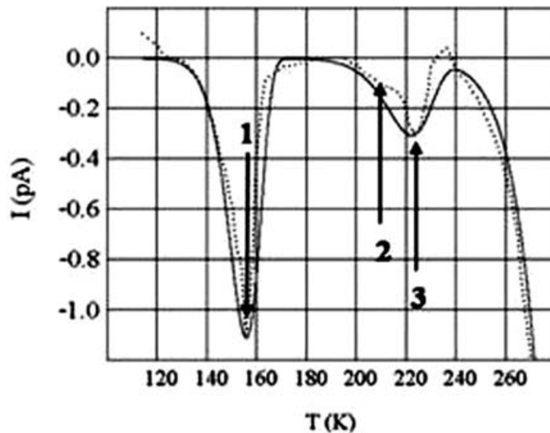


Fig. 8.15 Fitting of the OCS results for MQW structure: *solid line* – model; *dotted line* – experimental data [8]. Reprinted from Solid State Electronics, M. L. Ciurea, V. Iancu, and R. M. Mitroi, "Trapping Phenomena in Silicon-Based Nanocrystalline Semiconductors", **51**, 1328–1337, 2007, Copyright © 2007, with permission from Elsevier

Many power devices use silicon carbide. However, the SiC MOSFETs are faced with the shift of the threshold voltage and the lowering of the gate mobility, both induced by the traps located at the interface between the SiC and the gate oxide. This happens in the newly investigated high κ oxides, as well as in the classical SiO_2. The non-conventional method of *spin-dependent recombination* (SDR) was used to study the interface traps in an n-channel 6 H–SiC MOSFET with 50 nm thick SiO_2 gate oxide [37]. SDR method was first modeled by Lepine [38]. To investigate the deep trap levels (or recombination centers), one places the semiconductor device in a strong d.c. magnetic field. Then, both the conduction electrons (and holes) and the empty traps are spin-oriented. Under such conditions, the Pauli exclusion principle forbids the capture of the electron by the trap. When an RF magnetic field is applied orthogonal to the strong d.c. one, spin-flip occurs at the resonance condition (electron spin resonance – ESR) and the capture is abruptly increased. This is evidenced in capacitance–magnetic field (*C*–*B*) or resistance–magnetic field (*R*–*B*) measurements (see Figs. 8.16, 8.17). The superhyperfine peaks observed in Fig. 8.16 were interpreted in terms of interactions of ^{29}Si nuclei with a Si vacancy, allowing the identification of the observed deep trap level as a Si vacancy.

Another non-conventional method was proposed for the study of the carbon nanotubes [39]. A *low-energy electron point source* (LEEPS) microscope is used to image the shadow of a nanotube on an electron detector, as it can be seen from Fig. 8.18. The tip of the microscope emits electrons toward the detector. The nanotube is placed between them and therefore its shadow appears on the detector. If there is a local charge in the nanotube (e.g., charging a trap created by a twist of the nanotube), the shadow width will be modified (increased for a local negative charge and diminished for a positive one). The method allows the detection of one electron per 10 nm length. This way, charged traps can be individually detected and mapped.

The semiconductors from group IV, discussed up to now, have indirect band gap. The II–VI and III–V semiconductors have direct gap, allowing band-to-band radiative absorption and recombination. Therefore, they were studied especially for optoelectronic applications. Even in bulk semiconductors, to say

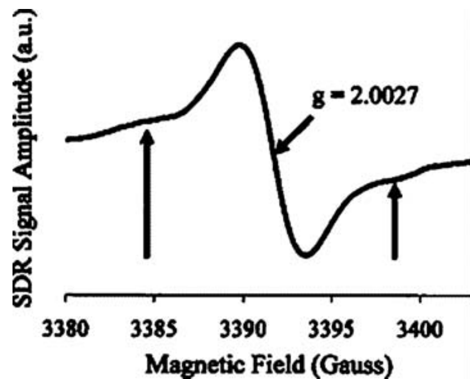


Fig. 8.16 *C*–*B* measurement of a 6 H–SiC MOSFET. Superhyperfine ^{29}Si peaks are marked by the arrows. The d.c. magnetic field is parallel with the *c* axis [37]. Reused with permission from D. J. Meyer, N. A. Bohna, P. M. Lenahan, and A. J. Lelis, Applied Physics Letters, **84**, 3406, 2004. Copyright © 2004, American Institute of Physics

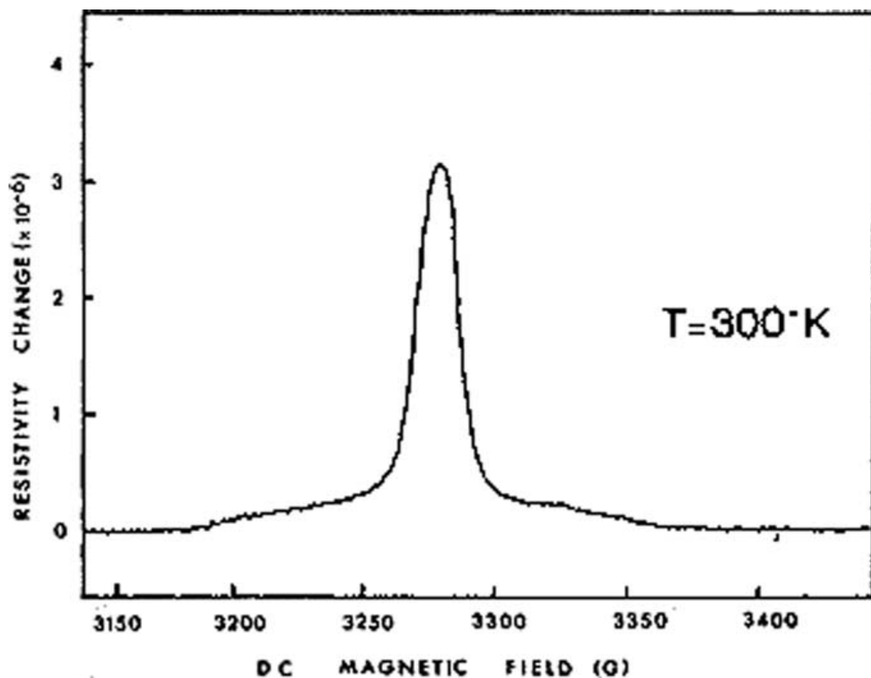


Fig. 8.17 *R–B* measurement of an n-type Si wafer [38]. Reprinted with permission from D. J. Lepine, Phys. Rev. B **6**, 436, 1972. Copyright © 1972 by the American Physical Society

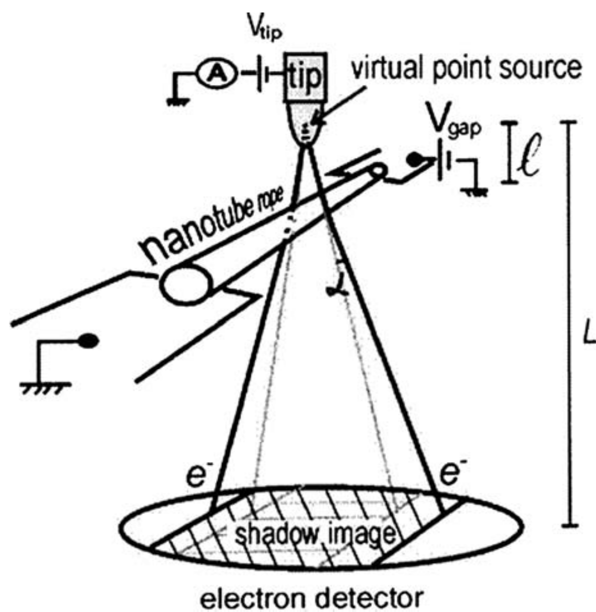


Fig. 8.18 LEEPS microscope set-up [39]. Reused with permission from P. S. Dorozhkin and Z.-C. Dong, Applied Physics Letters, **85**, 4490 2004. Copyright © 2004, American Institute of Physics

nothing about nanocrystals, the traps play a very important role in optical processes. This is why studying traps in nanocrystalline II–VI and III–V semiconductors is extremely important.

Traps in nanocrystalline $Cd_{0.8}Zn_{0.2}Te$ (II–VI semiconductor) thick films (40 μm thickness, with grain sizes of about 40 nm) were investigated by the analysis of the *time-of-flight* (TOF) technique in the transient photocurrent analysis [40]. To define TOF, a sample with at least one blocking electrode (that blocks one type of carriers) is necessary. Let us consider a sample with sandwich configuration where the top electrode is semitransparent, to allow the light to penetrate. If a strongly absorbed light pulse is applied, electron–hole pairs are generated close to the top electrode. The TOF is defined as the time needed by the carriers to reach the opposite electrode under the externally applied bias. This means that the TOF can be taken as the ratio between the sample thickness and the product of the carrier mobility with the applied electric field. The photocurrent produced by a short light pulse will abruptly drop after TOF. Using the Laplace transform and Tikhonov regularization methods, one can correlate the Laplace transform of the photocurrent with the trap density of states. This procedure requires very intricate calculations that are not presented here. The trap activation energy can then be estimated from the density of states maximum (considered as a function of energy).

The surface/interface traps in isolated nanocrystals can be evidenced by the study of *photoluminescence* (PL) *intensity fluctuations*. The isolated nanocrystals present strong PL intensity fluctuations, with long time intervals of darkness. This particular behavior is called *blinking*. The mechanism of electron–photon interaction which produces the blinking effect is correlated with the action of the surface traps [41]. The samples consist of CdS nanoparticles (5 nm diameter) immersed in a watery solution and spinned on a silica substrate. Two kinds of samples were studied: one of them containing bare nanoparticles and the other one with nanoparticles coated with ZnS. By using a confocal microscope, the PL of a single nanoparticle can be measured. An argon laser is used as excitation source.

The microscopy measurements show that the PL intensity oscillates between two states: an "on" state, with practically constant intensity, and an "off" state, with practically null intensity. The "on" time and "off" time are random and they can be described by probability distributions. The probability of measuring an "off" time value τ always follows an inverse power law, $P_{\text{off}}(\tau) \propto \tau^{-m}$, where m is an exponent experimentally determined. The case of the "on" time distribution is more intricate. The probability of measuring the "on" time for a coated nanoparticle also follows an inverse power law, $P^{\text{c}}_{\text{on}}(\tau) \propto \tau^{-n}$. On the contrary, the probability for bare nanoparticles is exponential, $P^{\text{b}}_{\text{on}}(\tau) \propto \exp(-a\tau)$.

The exponential behavior for the "on" times seems quite logical, as it agrees with the general relaxation law Eq. (8.1). To explain the power law for the "off" times, we have to consider the hopping of an electron between an excited nanocrystal and a trap. When the electron is trapped, the nanocrystal (charged with a hole) still absorbs light, but its de-excitation is non-radiative, through an

Auger recombination. The nanocrystal becomes bright when the electron hops back and radiatively recombines with the hole. The power law for the "on" times observed for the coated nanocrystals is explained considering that the hole from the coated nanocrystal cannot recombine as long as it is located in the core. Moreover, once the electron is trapped outside the core, the Coulomb blockade produced by the remaining hole will prevent a second ionization because the needed electrostatic energy is larger than the photon energy. If the hole is located on the shell, the core is photoactive, i.e., both absorption and radiative recombination can occur. By assuming two constant probabilities for these two locations, in the core and in the shell respectively, a power law can be deduced in which the exponent is dependent on the trap concentration. The healing of the surface traps determines a strong increase of the nanoparticle photoluminescence (PL) because of the reduction of the non-radiative recombination. This leads to several applications, like ultra-thin fluorescent dyes or biological markers (see also below).

The effects produced by the traps also appear in the absorption phenomena. If absorption measurements are compared with PL results, correlations regarding the traps arise. Thus, the study of the blinking PL and absorption spectra of isolated CdSe nanocrystals (3.9 nm diameter), bare, capped with octylamine (OA), or coated with CdS, gives information about the trapping phenomena [42]. From Fig. 8.19, one can observe that OA adsorption does not alter the absorption profile, while the coating induces a red shift of all the curve. The PL

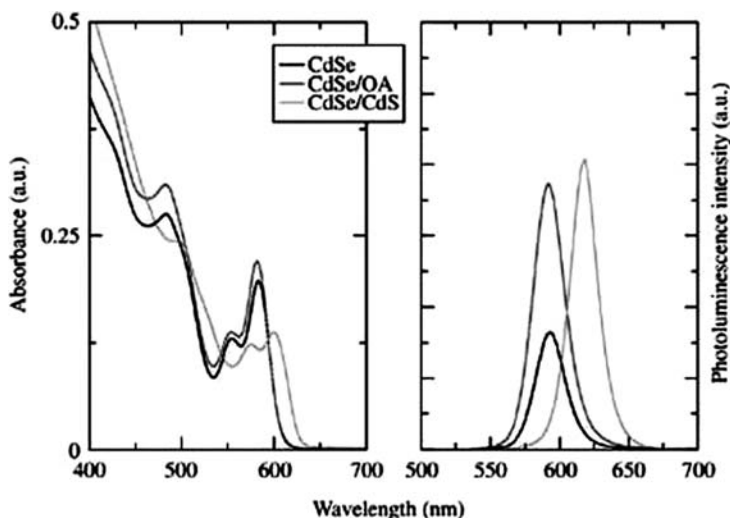


Fig. 8.19 Absorption (*left*) and integrated PL (*right*) spectra of the CdSe nanocrystals (3.9 nm core diameter), bare, capped, and coated [42]. Reproduced with permission from D. E. Gómez, J. van Embden, J. Jasieniak, T. A. Smith, and P. Mulvaney, "Blinking and Surface Chemistry of Single CdSe Nanocrystals" Small **2**, 204 (2006). Copyright © 2006 Wiley-VCH Verlag GmbH & Co. KGaA

curves (integrated over time) show a strong increase in intensity for both capped and coated nanocrystals. Again, the coating induces a red shift. The increase of PL intensity is due to the partial or total elimination of the surface traps (unsaturated dangling bonds) by surface passivation. At the same time, the red shifts that appear after the coating may be related to the QC effects. The increase of the diameter implies a decrease in the energy of the QC levels.

There are rather few data about III–V nanocrystals. The investigation of InAs nanocrystals, bare or coated with CdSe, by means of absorption and PL measurements is presented in Fig. 8.20 [43]. Both kinds of nanocrystals are capped with tri-n-octylphosphine (TOP) and dissolved in toluene. As in the previous case, one can observe a strong increase of the PL intensity, as well as a red shift, in the case of the coated nanocrystals. These results indicate the same interpretation: the disappearance of the InAs surface traps due to the coating, as well as a reduction of the peak energy following the increase of the nanodot diameter. Therefore, the healing of the surface traps by coating increases the PL signal by one order of magnitude, leading to important applications in biological experiments, where the nanocrystals are used as markers, as well as in the technology of optoelectronic devices.

The traps in $GaAs/In_{0.15}Ga_{0.85}As/GaAs$ quantum well layer were investigated by using another non-conventional method, namely by measuring its

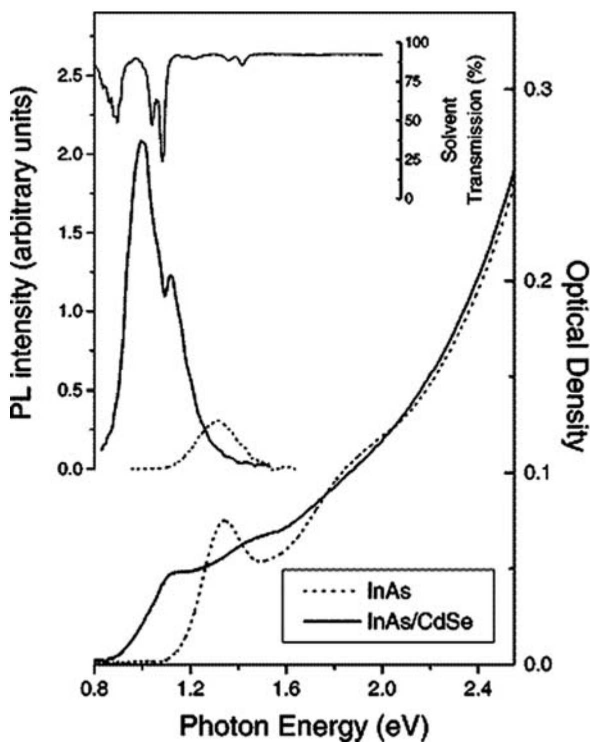


Fig. 8.20 InAs and InAs/CdSe nanocrystals absorption (optical density) and PL spectra [43]. Insert: toluene transmission spectrum. Reprinted with permission from C. McGinley, H. Borchert, D. V. Talapin, S. Adam, A. Lobo, A. R. B. de Castro, M. Haase, H. Weller, and T. Möller, Phys. Rev. B **69**, 045301, 2004. Copyright © 2004 by the American Physical Society

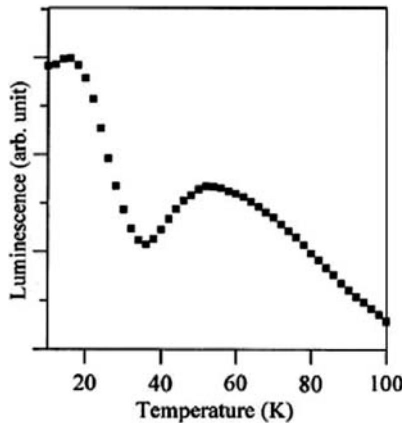


Fig. 8.21. Temperature dependence of the luminescence for a GaAs/$In_{0.15}Ga_{0.85}As$/GaAs quantum well [44]. Reprinted from J. Luminesc. **96** (2–4), M. Gal, L. V. Dao, E. Kraft, M. B. Johnston, C. Carmody, H. H. Tan, and C. Jagadish, "Thermally stimulated luminescence in ion-implanted GaAs", 287–293, Copyright © 2002, with permission from Elsevier

thermoluminescence (TL) as a function of temperature [44]. In order to do that, the sample is cooled down and illuminated for 10 min with strongly absorbed light. Then the excitation light is cut off and the sample is heated up at a constant rate, measuring the total luminescence (integrated over wavelengths) as a function of temperature.

Two maxima were observed (see Fig. 8.21). The maximum localized at 55 K can be eliminated by annealing the sample at 200°C, while the one at 15 K is not affected by the annealing process. The lower temperature maximum can be attributed to PL effects due to QC levels in the quantum well. The higher temperature maximum is due to TL effects produced by the thermally stimulated detrapping of the carriers from a shallow trap located in the GaAs barrier. The evaluation of the trap energy can be made by modeling both PL and TL effects [44]. This model for the interaction of the radiation with a substance does not present interest for this chapter.

As one can see from this paragraph, the non-conventional methods are strongly correlated with both the specific properties of the investigated systems and the positive or negative role played by the traps. On the other hand, it is better to use more than one method, in order to obtain information as complete as possible.

8.4 Summary and Concluding Remarks

The study of the trapping phenomena in nanocrystalline semiconductors proved that they take place mainly at the surface/interface of the nanocrystals. A special case is represented by the nanodots, where the trapping can also

appear on the QC levels. On the other hand, the Coulomb blockade raises several problems for the nanodots, like the partial filling of the traps, or the simultaneous discharge of two traps of opposite sign. The traps can improve the working parameters of some devices, or in some other cases, they can reduce the device reliability.

Most of the conventional methods used for the trap investigation in bulk semiconductors (like PICTS, TSC, TSDC, and OCS) are also applied to the nanocrystalline ones. For instance, the OCS method is a very sensitive one because the discharge current is a superlinear function of the trap concentrations. A thorough modeling allows the determination of trap parameters that are not directly measurable. However, some of the conventional methods (like DLTS) are not very suitable for nanocrystals. Several non-conventional methods with specific applicability were successfully introduced in the study of the traps. The SDR method determines the deep trap energies and presents the advantage to identify their nature, by means of the hyperfine interactions. Another non-conventional method that allows the estimation of the trap energy and concentration is the TOF technique for transient photocurrent measurements. On the other hand, the AFM and LEEPS microscopes permit the observation and the mapping of individual charged traps. Different investigation methods are complementary to each other and therefore ought to be used together to obtain full information about the traps.

The role of the traps in electrical processes was investigated mainly for the group IV semiconductors. As an example, the gate leakage current in MOSFETs is influenced by the traps located in the gate oxide or at its interface with the substrate. The trapped electrons contribute to the wear out of the oxide through the weakening of the Si–O bond. This leads to quasi-breakdowns (or soft breakdowns), i.e., the rather abrupt increase of the leakage current, reducing the reliability. The trap concentration is strongly increased by the presence of nanodots into the oxide. More than that, almost all the charge is stored on the nanodots if they are located at the edge of the oxide. If the nanodots are coated, the core/shell interface traps will act as a long-term electrical memory. At the same time, the apparent mobility of the carriers through the channel of a high κ MOSFET is sensibly reduced by the presence of traps in the dielectric. This is due to the increase in the effective threshold voltage and the decrease in the drain current, both induced by the charge accumulation in the traps located in the dielectric. On the other hand, the traps located in the vicinity of a nanodot influence its behavior by means of electrostatic interactions. Thus, the Coulomb blockade oscillations of the source-drain conductance in a single electron transistor are intensely modulated in amplitude and phase by the traps located near the nanodot. This can drastically affect the transistor behavior.

The trap contributions to the optical processes were studied mainly for II–VI and III–V semiconductors. The requirement of bright PL nanodots for biological markers and optoelectronic devices implies the healing of the surface traps. The passivation of the nanodot surface (CdSe, CdS, or InAs), either by capping or coating, proved itself a very good healing method. At the same time,

the shifts of the absorption and PL peaks for the coated dots could be related with the shift of the QC levels due to the diameter increase. A specific phenomenon for the II–VI nanodots, the blinking PL of a single nanodot, has been explained by the modeling of the oscillation of the carriers between the QC levels in a nanodot and the surface traps. On the other hand, the trapping on QC levels allows an efficient electroluminescence of Si nanodots subjected to an a.c. bias. This allowed the fabrication of a light source from a floating gate MOSFET, by inserting an array of quantum dots in the oxide layer.

The small number of atoms in a nanocrystalline semiconductor makes the contributions of the traps to different phenomena much more important than in bulk semiconductors. The trapping phenomena can be used to produce different devices, like floating gate memories, light-emitting transistors, and biological markers. On the other hand, the traps can reduce the reliability of the electronic devices and perturb their functioning. Therefore, their study is a stringent necessity.

Acknowledgments The work was partially supported from the CEEX-CERES 13/2006 Project in the frame of the First National Plan for Research and Development.

References

1. R. H. Bube, *Photoelectronic properties of semiconductors*. Cambridge University Press, pp. 1–70, 149–188 (1992).
2. S. M. Ryvkin, *Photoelectric effects in semiconductors*, Consultant Bureau, New York, pp. 1–19, 88–156 (1964).
3. D. A. Faux, J. R. Downes, and E. P. O'Reilly, J. Appl. Phys. **82**, 3754 (1997).
4. A. Benfilda, Proc. 1st Int. Workshop Semicond. Nanocryst. SEMINANO, Budapest 2005, **1**, 123 (2005).
5. S. Huang, and S. Oda, Appl. Phys. Lett. **87**, 173107 (2005).
6. J. Heitmann, F. Müller, L. X. Yi, M. Zacharias, D. Kovalev, and F. Eichhorn, Phys. Rev. B **69**, 195309 (2004).
7. M. L. Ciurea, V. S. Teodorescu, V. Iancu, and I. Balberg, Chem. Phys. Lett. **423**, 225 (2006).
8. M. L. Ciurea, V. Iancu, and R. M. Mitroi, Solid St. Electron. **51**, 1328 (2007).
9. E. Lusky, Y. Shacham-Diamand, A. Shappir, I. Bloom, and B. Eitan, Appl. Phys. Lett. **85**, 669 (2004).
10. S. Huang, S. Banerjee, and S. Oda, Mat. Res. Soc. Symp. Proc. **686**, A8.8.1 (2002).
11. S. Huang, S. Banerjee, R. T. Tung, and S. Oda, J. Appl. Phys. **93**, 576 (2003).
12. G. Bersuker, A. Korkin, Y. Jeon, and H. R. Huff, Appl. Phys. Lett. **80**, 832 (2002).
13. A. Neugroschel, L. Wang, and G. Bersuker, J. Appl. Phys. **96**, 388 (2004).
14. J. P. Campbell, P. M. Lenahan, A. T. Krishnan, and S. Krishnan, Appl. Phys. Lett. **87**, 204106 (2005).
15. D. V. Lang, J. Appl. Phys. **45**, 3023 (1974).
16. D. Cavalcoli, A. Cavallini, M. Rossi, and S. Pizzini, Fizika i Tehnika Poluprovodnikov **41**, 435 (2007).
17. G. L. Miller, IEEE Trans. Electron. Devices **ED-19**, 1103 (1972).
18. J. C. Balland, J. P. Zielinger, C. Noguet, and M. Tapiero, J. Phys. D. **19**, 57 (1986).
19. J. C. Balland, J. P. Zielinger, M. Tapiero, J. G. Gross, and C. Noguet, J. Phys. D. **19**, 71 (1986).

20. O. V. Brodovoy, V. A. Skryshevsky, and V. A. Brodovoy, Sol. St. Electron. **46**, 83 (2002).
21. I. S. Virt, M. Bester, M. Kuzma, and V. D. Popovych, Thin Solid Films **451–452**, 184 (2004).
22. T. Behnke, M. Doucet, N. Ghodbane, and A. Imhof, Nucl. Phys. B – Proc. Suppl. **125**, 263 (2002).
23. M. L. Ciurea, I. Baltog, M. Lazar, V. Iancu, S. Lazanu, and E. Pentia, Thin Solid Films **325**, 271 (1998).
24. P. Müller, Phys. Stat. Sol. A **23**, 165 (1974).
25. P. Müller, Phys. Stat. Sol. A **23**, 393 (1974).
26. T. Botila, and N. Croitoru, Phys. Stat. Sol. A. **19**, 357 (1973).
27. M. L. Ciurea, M. Draghici, S. Lazanu, V. Iancu, A. Nasiopoulou, V. Ioannou, and V. Tsakiri, Appl. Phys. Lett. **76**, 3067 (2000).
28. V. Iancu, M. L. Ciurea, and M. Draghici, J. Appl. Phys. **94**, 216 (2003).
29. J. Walters, G. I. Bourianoff, and H. A. Atwater, Nat. Mater. **4**, 143 (2005).
30. E. A. Boer, M. L. Brongersma, H. A. Atwater, R. C. Flagan, and L. D. Bell, Appl. Phys. Lett. **79**, 791 (2001).
31. M. Hofheinz, X. Jehl, M. Sanquer, G. Molas, M. Vinet, and S. Deleonibus, Eur. Phys. J. B **54**, 299 (2006).
32. M. L. Ciurea, V. Iancu, V. S. Teodorescu, L. C. Nistor, and M. G. Blanchin, J. Electrochem. Soc. **146**, 3516 (1999).
33. M. Draghici, M. Miu, V. Iancu, A. Nassiopoulou, I. Kleps, A. Angelescu, and M. L. Ciurea, Phys. Stat. Sol. A **182**, 239 (2000).
34. V. Ioannou-Sougleridis, A.G. Nassiopoulou, M. L. Ciurea, F. Bassani, and F. Arnaud d'Avitaya, Mater. Sci. Eng. C **15**, 45 (2001).
35. M. Draghici, L. Jdira, V. Iancu, V. Ioannou-Sougleridis, A. Nassiopoulou, and M. L. Ciurea, Proc. IEEE CN 02TH8618, Int. Semicond. Conf. CAS 2002, **1**, 119 (2002).
36. G. Bersuker, P. Zeitzoff, J. H. Sim, B. H. Lee, R. Choi, G. Brown, and C. D. Young, Appl. Phys. Lett. **87**, 042905 (2005).
37. D. J. Meyer, N. A. Bohna, P. M. Lenahan, and A. J. Lelis, Appl. Phys. Lett. **84**, 3406 (2004).
38. D. J. Lepine, Phys. Rev. B **6**, 436 (1972).
39. P. S. Dorozhkin and Z.-C. Dong, Appl. Phys. Lett. **85**, 4490 (2004).
40. K. H. Kim, K. N. Oh, and S. U. Kim, J. Kor. Phys. Soc. **41**, 471 (2002).
41. R. Verberk, A. M. van Oijen, and M. Orrit, Phys. Rev. B **66**, 233202 (2002).
42. D. E. Gómez, J. van Embden, J. Jasieniak, T. A. Smith, and P. Mulvaney, Small **2**, 204 (2006).
43. C. McGinley, H. Borchert, D. V. Talapin, S. Adam, A. Lobo, A. R. B. de Castro, M. Haase, H. Weller, and T. Möller, Phys. Rev. B **69**, 045301 (2004).
44. M. Gal, L. V. Dao, E. Kraft, M. B. Johnston, C. Carmody, H. H. Tan, and C. Jagadish, J. Luminesc. **96**, 287 (2002).

Chapter 9
Nanomechanics: Fundamentals and Application in NEMS Technology

Marcel Lucas, Tai-De Li, and Elisa Riedo

Abstract A nano-electromechanical system (NEMS) combines nanometer-sized actuators, sensors and electronic devices into a complex circuit. An intense effort has been made to develop versatile NEMS for the miniaturization of the existing devices and to design the new ones, with a wide range of applications in the field of electronics, chemistry and biology. All applications require a good understanding of the mechanical properties at the nanoscale and their influence on the other physical/chemical properties. In this chapter, the size dependence of the mechanical properties of nanostructures is discussed in detail and the influence of surface effects, defects and phase transitions is reviewed. The most commonly used techniques for studying the mechanical properties at the nanoscale are described and the potential applications of NEMS in biological/chemical sensing, data storage, telecommunications and electrical power generation are also presented.

9.1 Mechanical Properties at the Nanoscale

9.1.1 Introduction

Nanotechnology is a multidisciplinary field of science, which focuses on the manipulation, control and modification of matter at the scale of a nanometer (which is one billionth of a meter). A wide variety of nanostructures, such as nanowires, nanotubes and thin films, have been synthesized via top-down and bottom-up approaches [1]. Due to their small size and thus associated new properties, they are expected to contribute significantly to the miniaturization of existing technologies and to the development of new applications. Due to their stiffness, toughness and high aspect ratio, carbon nanotubes can be used as tips to increase the resolution of a scanning probe microscope. Nanowires may serve as interconnects in nanoelectronic, optoelectronic and spintronic devices.

E. Riedo
School of Physics, Georgia Institute of Technology, Atlanta, GA, USA
e-mail: elisa.riedo@physics.gatech.edu

A. Korkin, F. Rosei (eds.), *Nanoelectronics and Photonics*,
DOI: 10.1007/978-0-387-76499-3_9, © Springer Science+Business Media, LLC 2008

Epitaxial thin films are important for coatings and optoelectronic devices. A nano-electromechanical system (NEMS) combines actuators, sensors and electronic devices into a complex circuit, which is required in various biological/ chemical applications [2, 3].

All applications require a good understanding of the mechanical properties, namely elasticity and friction, at the nanoscale and their influence on the other physical properties (electronic, chemical, optical, etc.). A detailed study of the effect of mechanical deformation on the electron and heat transport properties of nanowires/nanotubes is required [4, 5, 6]. Residual stresses remain after the epitaxial growth of thin films, potentially affecting their optical properties, mechanical performances and resistance to heat [7, 8, 9]. The vibrational modes of carbon nanotubes shift in energy under uniaxial strain [10] and hydrostatic pressure [11]. Defects in carbon nanotubes induced by plastic deformation can enhance their chemical reactivity [12] and sensing capabilities [13]. The mechanical losses due to friction and adhesion have a negative impact on the operation of the silicon-based micro-electromechanical systems (MEMS) that involve sliding interfaces. Liquids have a solid-like behavior when confined in gaps of a few nanometers: for example, a 1 nm thick water film exhibits a viscosity which is orders of magnitude higher than the bulk viscosity [14].

Also, little is known about whether the classical models developed for materials at the microscopic scale still apply at the nanoscale. The continuum models describing the stretching of films or bending of wires or tubes must be thoroughly reviewed at the nanoscale, because of the significantly larger surface-to-volume ratios of the nanostructures. An understanding of the deformation and friction mechanisms in the mechanical nanodevices is important for the optimization of their performance.

So far, only a few experimental results are available, due to the technical challenges involved in preparing the samples and the lack of reliable methods to quantitatively measure the elasticity and sometimes the friction at the nanoscale. The problems are related to spatial and force resolution, instrument calibration, surface roughness and non-uniform chemistry (because at this scale each atom makes a difference). The size dependence of the elastic properties of nanostructures has been studied recently with the development of new techniques [15, 16] and different behaviors have been reported for nanostructures with various chemical composition. The Young's modulus of individual tungsten oxide (WO_3) nanowires was found to decrease significantly from the bulk value of about 300 to 100 GPa as the diameter increases from 16 to 30 nm [17]. The opposite trend was observed for GaN nanowires: the Young's modulus of a large nanowire of diameter 84 nm is consistent with the value in bulk GaN (300 GPa), but it decreases to about 220 GPa as the diameter decreases from 84 to 36 nm [18]. The elastic modulus of polystyrene films was also studied as a function of the film thickness. The elastic modulus of the thick films at penetration depths larger than 10 nm was close to the bulk value measured with a tensile test, but it decreased when the penetration depth was lower than 5 nm [19].

Some of these size-dependent mechanical properties can be explained by the behaviors observed on macroscopic samples. The increase of the elastic modulus in the polymeric nanofibers of diameters smaller than 70 nm was attributed to the improved alignment of the polymer chains with respect to the nanofiber axis [20], which is similar to macroscopic composite fibers [21]. The gold nanowires having a diameter as small as 40 nm have a plastic deformation mechanism which is dominated by the motion of dislocations, a behavior commonly observed in the macroscopic crystalline materials [22]. Similarly, the deformation mechanism of nanocrystalline gold films depends on its grain size. But even with a grain size as small as 100 nm, the macroscopic behavior, dominated by dislocation slip, is still observed [23]. Continuum models cannot fully describe the complexity of the mechanical behaviors at the nanoscale, since they neglect the atomic structure of the molecules or crystal. The elastic properties are predicted to be influenced significantly by the presence of grain boundaries in polycrystalline materials, free surfaces in high aspect ratio nanowires and point defects such as vacancies in the oxide nanostructures. For example, experimental results on nanocrystalline tungsten reveal a softening of the elastic constants at the nanoscale [24].

Apart from the size dependence of the elastic modulus, a wide range of new phenomena are also predicted in nanostructures by molecular dynamics simulations. The Young's modulus of nanotubes was observed to increase significantly as the diameter decreases below 1 nm [16, 25]. A theoretical study suggested that the Young's modulus depends not only on the diameter, but also on its chirality [26]. At high temperatures, they exhibit superplasticity, capable of sustaining an elongation of nearly 280% [27]. They are expected to undergo a series of reversible morphological changes, accompanied by the release of strain energy [28]. The high yield strength close to 400 MPa [29] and the superplastic extension above 5000% [30] of pure nanocrystalline copper suggest new deformation mechanisms dominated by grain boundaries. Finally, another possible way to release the strain energy at the nanoscale is a local phase transition. Under uniaxial tensile loading, ZnO nanowires can undergo a local and a reversible phase transition from the wurtzite structure, which is stable at room temperature, to a hexagonal structure [31].

9.1.2 Surface Effects

The surface-to-volume ratio becomes extremely high, once the dimensions of the sample decrease to a few tens of nanometers. The lower coordination of the surface atoms and the presence of surface charges can induce significant surface stresses that are well beyond the elastic regime [32, 33]. The charges present on the large polar surfaces of thin ZnO nanobelts can lead to spontaneous formation of helices, rings and coils [34].

The mechanical properties of thin films and nanostructures have been the subject of numerous theoretical studies. Attempts to extend the continuum models to the thinner films resulted in the systematic softening of the surfaces. By taking into account the atomic structure in the thickness direction, the Young's modulus of films that are thinner than 10 atomic layers was found to be 30% smaller than the bulk value [35]. Another continuum model, which distinguishes the surface elastic modulus and the bulk elastic modulus, not only predicts that the elastic constants of a nanoplate are inversely proportional to its thickness, but also shows that both softening and stiffening are possible in bars and plates thinner than 5 nm [36].

According to ab initio calculations on thin copper films, the softening or the stiffening is the result of competition between the low atomic coordination of the surface atoms and the charge distribution on the surface. The low coordination systematically softens the surface, and the magnitude of this softening depends on the direction along which the stress is applied with respect to the crystallographic faces. In contrast, a charge redistribution can lead to stiffening or further softening, depending on whether the electron density near the surface is increased or decreased, respectively [32]. Most of the theoretical studies show predominant surface effects only when the thickness of the film is of the order of a few nanometers. At a larger scale, non-linear bulk elastic properties can lead to orientation- and size-dependent behaviors [33].

Surface effects also explain the occurrence of peculiar phenomena at the nanoscale, such as shape memory and pseudoelasticity. Atomistic simulations of thin nickel and copper nanowires show that, during tensile loading, they undergo a reversible transition with the formation of defect-free twins and the modification of the crystallographic orientation of the side faces [37, 38].

Recent experimental results have revealed the major significance of surface effects, even at a scale of several hundreds of nanometers. Elastic constant measurements on tungsten and gold films (thickness around 250 nm) have indeed showed that both softening and stiffening can occur [24, 39]. Nanoindentation on thin polystyrene films also yielded a lower surface elastic modulus as compared with the bulk [19]. The size dependence of the elastic modulus of silver nanowires, lead nanowires and polypyrrole nanotubes was attributed to the surface tension effects. In this case, surface tension is defined as the energy required to elastically deform an already existing surface [40].

Other experimental data collected on nanowires were analyzed using a core–shell composite model, where the nanowire is considered as a composite material, comprising of a core which possesses bulk properties and a shell with different properties. In the case of the ZnO nanowires, the surface is stiffer than the core, due to the large compressive surface stresses [41]. As for the silver nanowires, the surface elastic modulus is affected by not only the surface stresses, but also the observed oxidation layer and the surface roughness [42].

9.1.3 Defects

The strength of a solid is a measure of its ability to resist plastic deformation. It is well established that in polycrystalline metals, the main plastic deformation mechanism is through the motion of dislocations. Therefore, crystallinity, defects and grain boundaries can significantly influence the bulk properties of materials. One way to strengthen the polycrystalline metals is to refine the grain size to the nanometer scale and thus introducing more grain boundaries. The tensile strength of the polycrystalline copper, having an average grain size of 400 nm and a high density of twin boundaries, is about 10 times higher than the bulk value. This remarkable result was attributed to the effective pinning of dislocations by the twin boundaries [43]. Reducing the dimensions of the samples, smaller than 1 μm, amplifies the effects of defects, as it significantly limits the multiplication and the motion of dislocations [44]. For example, the yield strength of a nickel alloy increases from 250 MPa to 2 GPa, as the diameter of the sample decreases from 20 to 0.5 μm [45]. Also, the spatial extent of the defects' influence on the mechanical properties of nanostructures might be larger than expected [46].

A size-dependent softening of the Young's modulus in tungsten oxide nanowires and carbon nanotubes was explained by the presence of defects. High-resolution transmission electron microscopy (TEM) revealed the presence of planar defects along the axis of the tungsten oxide nanowires with large diameters [17]. The density of defects in carbon nanotubes was quantified by Raman spectroscopy and a direct correlation was established between the Young's modulus and the concentration of defects [47]. Similarly, the porosity and the disorder have a negative effect on the mechanical properties of ZnO films [48] and polypyrrole nanotubes [49]. The elastic modulus of ZnO nanostructures was found to depend strongly on their width-to-thickness ratio, decreasing from about 100 to 10 GPa, as the width-to-thickness ratio increases from 1.2 to 10.3. This behavior was explained by a growth-direction-dependent aspect ratio and the presence of stacking faults in the nanobelts grown along particular directions (Fig. 9.1) [50].

Numerous discontinuities were observed in the force–indentation curves collected from gold and zinc oxide crystals. These are the evidence of the formation, multiplication and slip of dislocations [51]. After nanoindentation, the TEM images of the cross-section of a zinc oxide crystal showed no cracks, but dislocations and slip planes along particular crystallographic planes were observed [52]. A comparison between the force–indentation curves collected on an atomically flat gold surface and a gold surface with surface atomic steps revealed that the steps act as nucleation sites for dislocations [53].

However, defects can also have a positive effect on the mechanical properties of materials. The implantation of ions (point defects) reduces dislocation slip, leading to an increase in hardness of MgO crystals on all crystallographic surfaces and also reduced pile-up around the indent [54]. Collisions between carbon nanotube bundles and high-energy electrons in a TEM lead to the

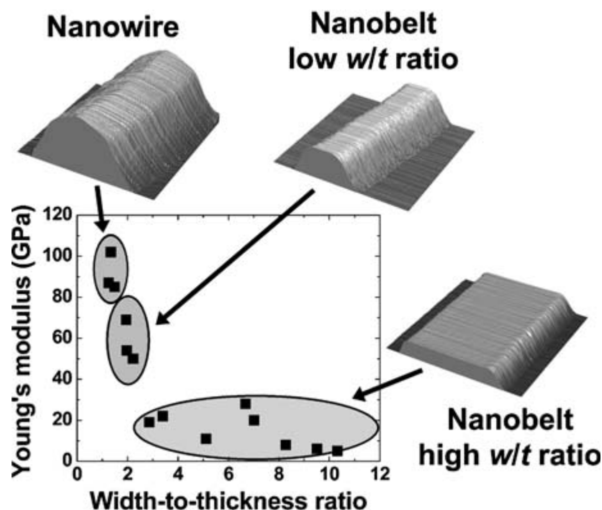


Fig. 9.1 Young's modulus of ZnO nanostructures as a function of width-to-thickness ratio w/t. (Reproduced with permission from [50]. Copyright 2007, American Chemical Society)

formation of vacancies along the nanotube sidewall. Atomic rearrangements with the neighboring tubes in a bundle or other adsorbates create cross-links between the nanotubes, limiting sliding and resulting in an increase of the bending modulus by a factor of 30 [55]. The silver nanowires produced via a chemical process have five internal twin boundaries over their entire length, hence giving them a pentagonal structure. The presence of these twin boundaries is expected to stabilize their structure and prevent the propagation of dislocations during tensile loading [56]. The elimination of the five twin boundaries by thermal treatment does not affect their Young's modulus, but makes them more ductile [57].

Early attempts to develop a model to study the plastic deformation via the formation and motion of dislocations revealed the importance of the strain gradient effects on the mechanical properties of crystalline solids, particularly in the nanoindentation tests [58]. The classical theories, which only include strain effects, contain no length scale and therefore cannot explain the depth dependence of hardness at depths that are below 1 μm. Strain gradients in the indent area result in higher hardness, because of the generation of geometrically necessary dislocations. The magnitude of these strain gradient effects is amplified at low indent depths. Improvements made on a strain gradient theory of plasticity introduced characteristic length scales related to the dislocation source, the motion and the interactions between dislocations that are measurable material quantities. The theory successfully described the size-dependence of the hardness for polycrystalline and single-crystal copper, as well as for single-crystal silver along two different crystallographic directions [59]. When the dimensions of the sample, such as the nanowire diameter or the grain size in polycrystalline metals, fall below these characteristic length scales, dislocation motion is limited and other deformation mechanisms become dominant. For

example, a pure nanocrystalline copper with an average grain size of 80 nm exhibits a near-perfect elastoplastic behavior and also deforms homogeneously without any neck formation. This behavior is attributed to a deformation mechanism that is dominated by the diffusion of grain boundaries [29].

9.1.4 Phase Transitions

The mechanical properties of a single crystal depend on its crystal structure and are highly anisotropic. For example, the elastic modulus measured by nanoindentation on the (111) face of a gold single crystal is higher than the one measured on the (110) and (001) surfaces [60]. Ab initio calculations on the tensile loading of thin ZnO nanobelts show that this anisotropy remains and is even amplified when the lateral dimensions of the nanobelts fall below 3 nm [61]. In macroscopic polymeric samples, crystallinity has a significant effect on their mechanical properties, since the crystalline and amorphous regions have very different characteristics. Stress-induced crystallization of elastomers has been observed experimentally, and it leads to an enhanced tensile strength and resistance to crack propagation [62]. A similar phenomenon is also possible in very thin nanostructures, opening another way to release the strain energy introduced during the tensile loading and hence possibly resulting in an apparently low Young's modulus. At ambient conditions, ZnO has a wurtzite structure, but at a pressure close to 9 GPa, a phase transition to a rocksalt structure was observed. The molecular dynamics simulations show that ZnO nanowires undergo a reversible phase transition under tensile loading at a critical strain, accompanied by a sudden stress drop in the stress–strain curve [31]. Such a phase transition is facilitated by surface stresses and can potentially affect the piezoelectric properties, the electronic and thermal conductivities of the nanostructure [63]. The occurrence of a phase transition depends on the direction of the applied strain with respect to particular crystallographic planes. The Young's modulus of ZnO nanoplates exhibiting wide (0001) surfaces is expected to vary discontinuously as a function of thickness, because of phase transformations from a wurtzite to a graphitic structure [64]. The experimental evidence of such phase transitions has been the subject of nanoindentation and TEM studies. Discontinuities in the force–displacement curves could be the signature of a transformation, but so far, no additional TEM data support this conclusion. It was also suggested that the appearance and the number of discontinuities depend on the crystal structure and also on the crystallographic surface probed by nanoindentation [51].

9.2 Methods

Due to the small size of nanostructures, new characterization techniques had to be developed and they must combine an extremely high spatial resolution and high force sensitivity. The forces that are required to deform a nanostructure

are of the order of a few nanonewtons, and even less for more fragile samples, such as cell membranes or proteins. A suitable characterization method must also include an imaging capability, since the mechanical properties are closely related to the dimensions, morphology and crystallographic structure of the sample. So far, most of the experimental data are obtained from atomic force microscopy (AFM) and electron microscopy. When scanning conditions are optimal, AFM attains imaging with atomic resolution and sub-nanonewton force resolution. TEM also offers atomic spatial resolution and the capability to characterize the structure of single crystals or detect defects. In the following section, AFM- and TEM-based techniques used to study the mechanical properties at the nanoscale are described. Other techniques, such as optical tweezers and spectroscopy, are also discussed.

9.2.1 Scanning Probe-Based Methods

Atomic force microscopy (AFM) is an ideal tool for investigating the mechanical properties due to its ability to directly measure forces between the tip and the sample with a nanonewton resolution. The force components normal and parallel to the substrate are measured by monitoring the deflection of the cantilever, scanned over the sample in contact. A laser beam, reflected from the back of the AFM cantilever to a four-quadrant photodetector, provides the force measurement and a feedback mechanism for high-resolution imaging. In the following section, some of the AFM-based techniques, to study the mechanical properties at the nanoscale, are reviewed.

9.2.1.1 Force-Displacement Curves

The simplest way to study the mechanical properties of nanomaterials is to collect a force–displacement or a force–indentation curve. The AFM can measure forces applied by the tip to the sample as a function of the displacement of the scanner supporting the sample. AFM nanoindentation has been extensively used to study the elastic modulus and the hardness of polymer films [19] and single crystals [51, 54, 60]. The high force and displacement resolution of the AFM enables the indentation of a sample to depths as low as several nanometers. However, for nanostructures with a thickness of only a few nanometers, indentation results are influenced by the stiff substrate. To eliminate the substrate influence, nanostructures can be deposited off the edge of the substrate or over a trench (Fig. 9.2). For example, Wong et al. used an AFM to image individual and structurally isolated silicon carbide nanorods and carbon nanotubes that were pinned down at one end on a molybdenum disulfide surface by the deposition of silicon oxide pads, thus leaving the other end free to deform [65]. Then an AFM tip approaching from the side would bend the protruding part of the nanostructure. The lateral deflection of the cantilever is measured as the tip is scanned along a direction perpendicular to the

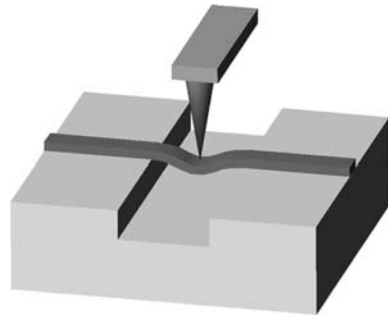


Fig. 9.2 Three-point bending test of a suspended nanostructure. An AFM tip applies a bending load in the middle of the suspended section of the nanostructure

nanostructure axis, yielding a force–bending deformation curve. Another possibility is to deposit nanostructures over a porous membrane where the suspended section is free to deform. Salvetat et al. [66] used this method to investigate the modulus of carbon nanotubes deposited on an alumina ultrafiltration membrane. An AFM tip would approach from above the nanostructure and will come in contact with it in the middle of the suspended section. The force applied by the tip to the sample is then measured as a function of the deformation.

The force vs. bending curve is then analyzed with the elastic beam-bending theory, which describes the mechanical properties of beam-like materials, such as nanorods [65], nanotubes [66], nanowires [22, 67, 68, 69] and nanobelts [70]. The elastic modulus can be extracted from the beam theory, which takes into account the geometry of the nanostructure and also the boundary conditions of the system. The mechanical response of the nanostructure depends on whether it is considered as free at both ends, clamped at only one end or clamped at both ends. It is reasonable to consider the double-clamped beam model when a nanostructure is clamped by metallic or oxide pads at both ends [22, 69], even if slippage is possible with a poor interface. However, the double-clamped beam model was also used for carbon nanotubes that were simply deposited on a porous membrane without the deposition of pads. It was argued that the adhesion force between the carbon nanotube and the alumina membrane is much larger than the force applied to indent the nanotube. Setting inappropriate boundary conditions could yield elastic modulus values that differ by a factor of 4 [70].

Aside from the boundary conditions, other potential sources of errors include the inaccurate measurements of the nanostructure dimensions and the inaccurate positioning of the AFM tip. In the case of a nanostructure suspended over a trench, the load must be applied at the middle point of the suspended section. Instead of collecting a single force–deformation curve at the middle point, Mai et al. proposed to measure the deformation profile of the entire suspended section by collecting AFM images at different set points. The deformation profile can then be fitted with the beam model and suitable boundary conditions, hence eliminating the need to position the tip over the middle point and the uncertainty over the boundary conditions [68, 70]. The

force–displacement curve methods can be applied to a wide variety of nanostructures, since they are suitable for soft samples, such as cells [71], and also to study chemically specific interactions between molecules [72]. However, a large number of force–displacement curves are required to reduce errors, which is time-consuming and thus limits the number of nanostructures that can be studied.

9.2.1.2 Lateral Force Imaging

The samples have to be prepared carefully for the acquisition of force–displacement curves. For nanostructures deposited over a trench or off the substrate edge, the nanostructure long axis should be perpendicular to the substrate edge. Most of the nanostructures are in bundles (carbon nanotubes) or entangled (ZnO nanobelts) after their synthesis. It may be difficult to isolate them and align them, due to their brittleness or the lack of appropriate tools. This is the case for arrays of vertically aligned ZnO nanowires attached at one end on a sapphire substrate. Another technique based on an AFM was proposed to measure their elastic modulus without damaging them [73]. When an AFM tip is scanning parallel to the substrate, the top end of vertically aligned nanowires experiences a lateral force f perpendicular to the nanowire (Fig. 9.3). The displacement of the nanowires can be expressed under the small deflection approximation by the elastic beam model as

$$EI\frac{\mathrm{d}^4x}{\mathrm{d}y^4} = (f_0 + f)\delta(y - L) \tag{9.1}$$

where f_0 is the component of the friction force between the tip and the nanowire along the scanning direction; E and I are the elastic modulus and momentum of

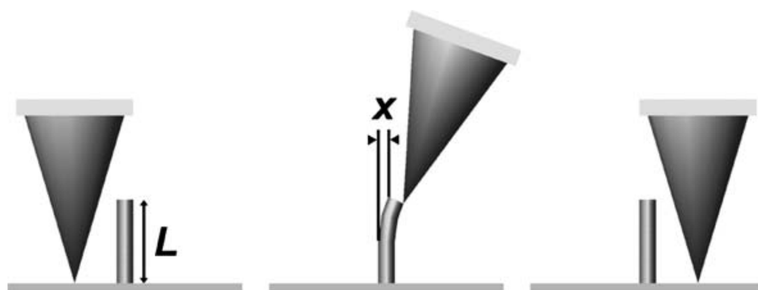


Fig. 9.3 Measuring the elastic modulus of vertical nanowires from their lateral bending with an AFM tip. As the tip is scanned from left to right, the tip comes into contact with a nanowire. (**a**) Before and (**c**) after contact, only a small lateral signal is detected. (**b**) When the tip is in contact with the nanowire, the scanner retracts and the tip is deflected laterally, resulting in a large lateral signal. L is the nanowire length and x the lateral displacement, perpendicular to the nanowire axis. (Reproduced with permission from [73]. Copyright 2005, American Chemical Society)

inertia of the nanowire, respectively. x is the lateral displacement perpendicular to the nanowires, y is the height from the fixed end (root) of the nanowire to the point where the lateral force is applied, which is approximately the tip of the nanowire ($y = L$, the length of the nanowire), and the contact is assumed to be a point. f_0 is much smaller than the bending force f especially when the scanning speed is low, and therefore can be neglected. The applied lateral force f is expressed as

$$f = 3EI\frac{x}{L^3} \tag{9.2}$$

From Hooke's law, the spring constant is $K = f/x$; thus, the elastic modulus can be expressed as a function of K, L and I: $E = KL^3/3I$. A ZnO nanowire grown along [0001] usually has a hexagonal cross-section with a side length a (a is considered as the radius of the nanowire), and in this case the momentum of inertia is $I = (5(3^{1/2})/16)a^4$. SEM and TEM images indicate that the aligned ZnO nanowires have a uniform diameter of 45 nm and the heights vary from 200 to 800 nm. Thus, the elastic modulus is given by

$$E = \frac{16L^3K}{15(3^{1/2})a^4} \tag{9.3}$$

Using this technique, the elastic modulus of 15 nanowires was measured with a single lateral force image. The lengths of the nanowires vary from 170 to 680 nm and the corresponding elastic modulus ranges from 15 to 47 GPa, with an average value of 29±8 GPa.

9.2.1.3 Modulated Nanoindentation

Bending tests usually deform the nanostructures beyond their linear elastic regime. Therefore measuring the elastic properties of nanostructures remains a technical challenge. For example, to measure the radial elastic modulus of carbon nanotubes (diameter of several nanometers) with force vs. deformation curves, the AFM would have to measure forces of a few nanonewtons against displacements of a few angstroms. An alternative method was proposed to increase the force and displacement resolution: the modulated nanoindentation method, which was proven to be a powerful technique to measure the radial elastic modulus of carbon nanotubes [25] and study the structure-dependent elastic modulus of ZnO nanobelts [50, 74]. The modulated nanoindentation method consists in indenting a sample (simply deposited on a stiff substrate such as silicon) with an AFM tip while oscillating the sample in the normal direction with a piezoelectric scanner excited with an ac signal of fixed amplitude and frequency (Fig. 9.4). The displacement of the piezoelectric scanner d_{tot} is the sum of contact deformation (indent depth) and cantilever bending. The amplitude of the oscillations d_{tot} is typically a few angstroms. Instead of

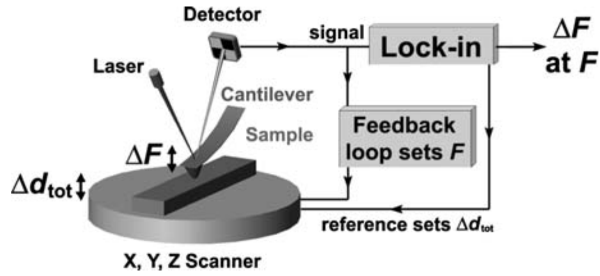


Fig. 9.4 Experimental setup for the modulated nanoindentation method. (Reproduced with permission from [50]. Copyright 2007, American Chemical Society)

measuring the normal force as a function of indentation depth, $F(z)$, the slope $\frac{\mathrm{d}F}{\mathrm{d}z}$ is measured around a fixed force set point F_0 using a lock-in amplifier. In practice, the force variations ΔF are recorded at different values of F_0 and at a fixed Δd_{tot}.

The tip–sample system is then modeled as two springs in series and its total stiffness k_{total} is given by

$$\frac{\partial F}{\partial d_{\text{tot}}} = k_{\text{total}} = \left(\frac{1}{k_{\text{lever}}} + \frac{1}{k_{\text{contact}}}\right)^{-1} \tag{9.4}$$

where k_{lever} and k_{contact} are the stiffness of the AFM cantilever and the tip–sample contact, respectively. Analytical expressions of k_{contact} are obtained by considering the geometry of the tip–sample system. For example, in the case of carbon nanotubes, the contact is similar to a sphere (tip) indenting a cylinder, in which case

$$k_{\text{contact}} = \beta\left(\frac{R(F_0 + F_{\text{adh}})}{\tilde{K}^2}\right)^{\frac{1}{3}} \tag{9.5}$$

with $\frac{1}{R} = \frac{1}{R_{\text{tip}}} + \frac{1}{2R_{\text{ext}}}$, R_{tip} the tip radius, R_{ext} the external nanotube radius, F_{adh} the adhesion force between the tip and the sample and $\tilde{K} = \frac{3}{4}\left(\frac{1-\nu_1^2}{E_1} + \frac{1-\nu_2^2}{E_2}\right)$, where $\nu_{1,2}$ and $E_{1,2}$ are the Poisson's ratios and Young's moduli of the tip and the nanotube (along the radial direction), respectively. β is a coefficient that takes into account the geometrical aspect of the contact area [25]. The Young's modulus E_2 of the nanotube is the only fitting parameter for the k_{contact} vs. F_0 curve. Applying this method, the radial Young's modulus of carbon nanotubes was found to increase sharply as the external radius decreases when R_{ext} is smaller than 4 nm and remains constant at 36 GPa when R_{ext} is larger than 4 nm (Fig. 9.5).

9.2.1.4 Contact Stiffness Mapping

Acoustic microscopy is a non-destructive technique developed to study the elastic properties of macroscopic samples and to detect the presence of defects.

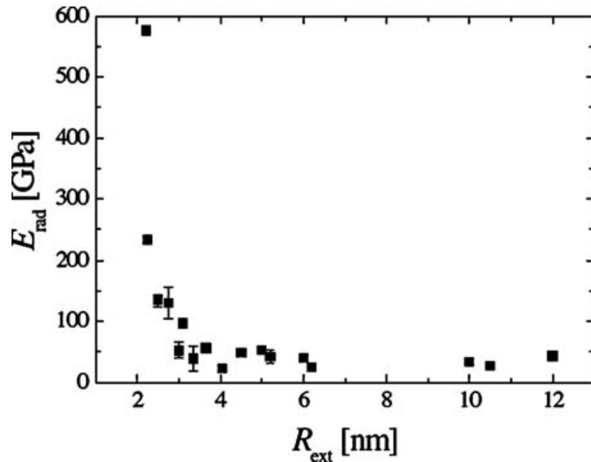


Fig. 9.5 Experimental values of the radial Young's modulus E_{rad} of carbon nanotubes as a function of the external radius R_{ext} obtained from normal modulated nanoindentation. (Reprinted with permission from [25]. Copyright 2005, the American Physical Society)

However, its spatial resolution is limited by the wavelength of the excitation probe, which is of the order of a micron. This resolution limit can be overcome by the detection of ultrasonic vibrations (MHz to GHz range) with an AFM tip. In an ultrasonic force microscope (UFM), an ultrasonic excitation of a few MHz is applied to a piezoelectric actuator placed under the sample. This high-frequency wave is modulated by another signal, either triangular or trapezoidal, at a frequency below a few kHz. When the AFM tip is scanned in contact over the sample, the amplitude of the cantilever vibration varies because of the non-linearity of the tip–sample interactions or contact stiffness. By comparing the amplitude variations on the sample to the ones on well-known materials, the stiffness of the sample can be extracted. For example, UFM was applied to the investigation of lattice defects in highly oriented pyrolytic graphite [75]. In contrast to TEM, thin samples are not required with UFM and topographic features such as surface steps and dislocations can also be detected, when the sample is not excited by the ultrasonic wave.

In contrast to UFM, acoustic force atomic microscopy (AFAM) monitors the resonance frequencies of the cantilever to extract the local contact stiffness. The cantilever is excited at one of its contact resonance frequencies in the MHz range. Usually, imaging with the cantilever vibrating in a high-frequency mode increases the sensitivity to contact stiffness variations. When the AFM tip is scanned over the sample, variations in the local elasticity are detected by a resonance frequency shift relative to the value of the free cantilever. By measuring the resonance frequency and its shift with a reference sample, the elasticity of the sample can be determined [76]. A combined study of the elastic modulus of SnO_2 nanobelts with UFM and AFAM showed an excellent quantitative agreement, and the measured elastic modulus value was consistent with the value obtained from nanoindentations [77].

9.2.2 Electron Microscopy

Electron microscopy, in particular TEM, is a powerful imaging technique, which can measure deformations or displacements with atomic resolution. It can be combined with electron diffraction to study the crystallographic orientation or lattice spacing of single crystals or electron energy-loss spectroscopy for a local chemical analysis. In the following section, the methods to extract the elastic modulus or the tensile strength of nanostructures with a TEM are reviewed.

9.2.2.1 Mechanical Resonance

The intrinsic thermal vibrations of multiwalled carbon nanotubes, clamped at one end on a substrate, were used to measure their Young's modulus. TEM images allowed the precise measurements of the nanotube diameter, length and vibration amplitude of its free end. The vibration amplitude was measured for different temperatures between the room temperature and 800°C. The Young's modulus was then extracted by analyzing the data with the Bernoulli–Euler theory of elastic beams. An average value of 1.8 TPa for the Young's modulus was found over 11 samples, with an external radius ranging from 6 to 25 nm [78].

Instead of relying on the measurement of small vibration amplitudes, a more precise method based on the electromechanical resonance of nanostructures was proposed. The mechanical resonance method is a non-destructive method, where a vibration mode of the nanostructure is excited by a periodic electric field between the electrode supporting the nanostructure and a counterelectrode (Fig. 9.6). An ac voltage is applied between the two electrodes, and the vibration amplitude is measured from TEM images while the frequency of the

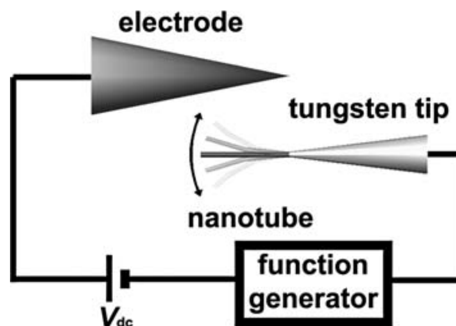


Fig. 9.6 Experimental setup used to study the mechanical resonance of nanostructures in an electron microscope. The nanostructure vibrates when an ac voltage is applied between the probe and the electrode. A dc offset is applied to increase the amplitude of the vibrations. (Reused with permission from [79]. Copyright 2005, American Institute of Physics)

excitation varies. For nanotubes, the resonant frequencies ν_j are given by the Bernoulli–Euler theory of cantilevered elastic beams:

$$\nu_j = \frac{\beta_j^2}{8\pi} \frac{1}{L^2} \sqrt{(D^2 + D_i^2)} \sqrt{\frac{E}{\rho}} \tag{9.6}$$

where D is the outer diameter, D_i the inner diameter, L the length, E the elastic modulus, ρ the density and β_j a constant for the jth harmonic: $\beta_1 = 1.875$, $\beta_2 = 4.694$. This technique was used to study the mechanical properties of nanowires [17, 18] and nanotubes [16, 79].

9.2.2.2 In Situ Tensile or Bending Test

Another method to measure in situ forces applied to nanostructures is to use AFM cantilevers as force transducers. After calibrating the cantilever spring constant, the bending of the cantilever measured with TEM images enables the measurement of the force applied by the tip on a nanostructure. Enomoto et al. [47] integrated AFM cantilevers in a stage that can fit inside a TEM and used them to measure the Young's modulus of carbon nanotubes fixed at one end at the tip of an aluminum wire (Fig. 9.7). A silicon cantilever, mounted on a piezoelectric XYZ stage, approaches and applies a bending load on an individual nanotube. The bending load is applied in incremental steps and the deflection of the cantilever, the bending of the nanotube and the position of the contact point are obtained by acquiring TEM images at each step. The nanotube elastic modulus is then extracted by analyzing the force–displacement curve with the elastic beam model.

Another important mechanical property of 1D nanostructures is the tensile strength, which is accessible only with a tensile test. Tensile tests on individual multiwalled carbon nanotubes were performed inside a scanning electron microscope. The ends of the nanotube were attached to two different AFM

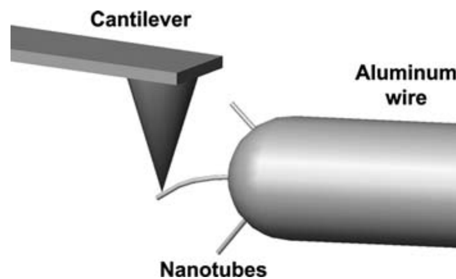


Fig. 9.7 Schematic of a nanotube bending experiment. The nanotubes are attached to the tip of an aluminum wire. An AFM cantilever applies a bending load to an individual nanotube. (Reused with permission from [47]. Copyright 2006, American Institute of Physics)

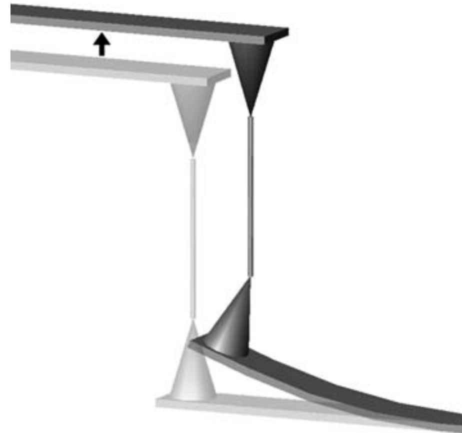


Fig. 9.8 Schematic of a tensile-loading experiment. An individual nanotube is attached to a stiff cantilever (*top*) and a soft cantilever (*bottom*). The bending of the soft cantilever is monitored as the stiff cantilever is driven upward by a piezoelectric motor. (From [80]. Reproduced with permission from AAAS)

cantilevers, one stiff with a high spring constant and another soft with a low spring constant, by using the electron beam to deposit some solid carbonaceous material. As the stiff cantilever is driven away from the soft cantilever by a piezoelectric actuator, the soft cantilever is bent and acts as the force transducer. By acquiring SEM images continuously, during the tensile test, the nanotube strain, the bending of the soft cantilever and thus the force applied to the nanotube can be obtained simultaneously (Fig. 9.8). The tensile strength measured for the outer layer of the nanotube ranged from 11 to 63 GPa for 19 samples. The stress–strain curves also yielded Young's modulus values between 270 and 950 GPa [80].

The main difficulty during the test is to control the alignment of the nanotube with respect to the cantilevers. The axis of the applied load must be maintained aligned along the nanotube axis. Any misalignment leads to a lower measured load and a higher measured strain. The Young's modulus of crystalline boron nanowires was measured with the mechanical resonance method and tensile tests in a SEM, yielding comparable values between 300 and 400 GPa [81].

9.2.3 Optical Methods

TEM-based methods have been widely used to characterize nanostructures, because of the high-resolution imaging capability, but they also have some limitations. The mechanical resonance method is difficult to apply on a soft sample or one with an irregular shape. Also, the sample must be thin, with sufficient electrical conductivity. The characterization of polymer nanofibers or biomolecules is not possible due to their poor electrical conductivity. Therefore, other non-destructive methods have been developed using laser beams and

optical spectroscopy. The main advantage of optical methods is their ability to manipulate and probe the nanomaterial without any physical contact.

The development of optical tweezers was a major advance in the manipulation of nanomaterials and biomolecules. Optical trapping with a single laser beam was demonstrated on a wide range of particles down to a diameter of 25 nm in water [82]. When a laser beam is strongly focused, for example by a high numerical aperture microscope objective, an intense laser intensity gradient is generated along the incident laser beam path. When a neutral dielectric particle is in the beam path, the light scattered by the particle results in a force, which is proportional to the intensity gradient and directed toward the beam focus where the intensity is highest. The optical trap is most effective when the particle size is comparable to the laser wavelength. In practice, an external force applied to the particle can be determined from the resulting displacement off the laser beam focus if the laser wavelength, intensity and the refractive index of the medium are known. When the particle displacement is small, the particle acts as a spring, which follows Hooke's law. Small particles inside this optical trap can then be manipulated and used as the force transducers. This technique is widely used for the study of the elasticity of biomolecules attached to microparticles. Force–distance curves were collected on double-stranded and single-stranded DNA. Using optical tweezers, the force required to separate two strands of DNA or to pack DNA inside a viral capsid was determined with a resolution of several piconewtons [83].

Another useful optical method relies on Brillouin light scattering, which is based on the inelastic interaction between photons and phonons. Brillouin spectra are interpreted on the basis of Lamb's theory to extract elastic properties, such as Young's modulus, shear modulus and Poisson's ratio. Brillouin scattering on a single isolated silica sphere, with a diameter as low as 260 nm, yielded a Young's modulus of 33 GPa and a Poisson's ratio of 0.18 [84]. A polarization study can potentially determine the anisotropic elastic constants of single crystals.

9.3 Applications

9.3.1 Sensors

AFM cantilevers were first developed for imaging purposes only, but their high sensitivity combined with a low mass-production cost makes them ideal for sensing applications. Small variations of the physical and/or chemical properties at the surface lead to the bending of the cantilever, which can be measured by a laser beam deflection, capacitive sensors or a piezoelectric bimorph. An asymmetric coating of the cantilever with a chemically functionalized layer favors the adsorption of molecules on the functionalized surface that induces, in most cases, a bending of the cantilever due to electrostatic repulsions or steric effects (Fig. 9.9a). Based on this principle, two forms of a prostate-specific

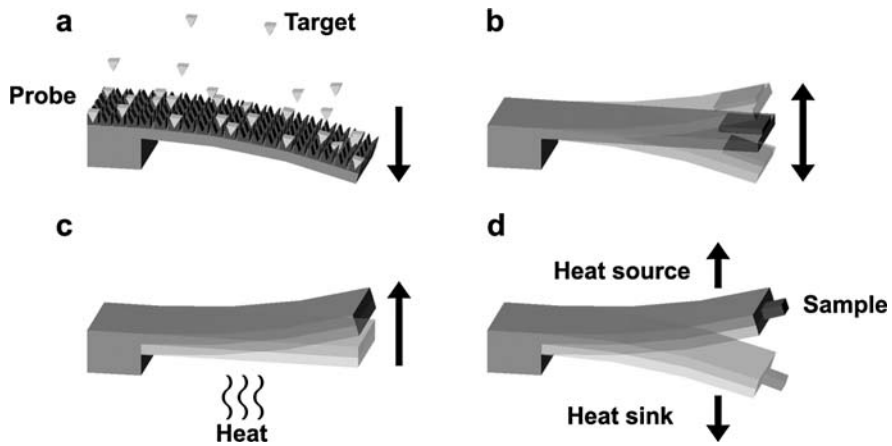


Fig. 9.9 Operating principles of a cantilever-based (**a**) chemical sensor, (**b**) mass sensor, (**c**) heat sensor and (**d**) calorimeter. (Reproduced with permission from [88]. Copyright 2002, IOP Publishing Ltd)

antigen, a marker for detection of prostate cancer, were detected using cantilevers coated with specific antibodies. An antigen concentration as low as 0.2 ng/ml was detected, which is 20 times lower than the threshold of 4 ng/ml required for clinical tests [85].

The mechanical resonance of oscillating cantilevers can also be exploited to measure the mass of small adsorbates. When the cantilever is driven at its resonance frequency, a variation of the cantilever mass M due to the adsorption or desorption of molecules induces a shift in the resonance frequency ν, similar to the concept used for macroscopic quartz oscillators (Fig. 9.9b). The relationship between the resonant frequency and the mass of the cantilever is given by

$$\nu = \frac{1}{2\pi}\sqrt{\frac{k}{M}} \tag{9.7}$$

where k is the spring constant of the cantilever, which is obtained from the resonance frequency of the unloaded cantilever. The adsorption of *Escherichia coli* on a cantilever coated with an antibody layer was studied quantitatively and the detection of only 16 *E. coli* cells, corresponding to a mass of about 6 pg, was demonstrated [2]. Using cantilevers with dimensions smaller than a micron, attogram sensitivity in ambient conditions and even zeptogram sensitivity at cryogenic temperatures in ultrahigh vacuum were achieved [86]. In addition, cantilevers present multiple resonance modes that can behave differently with varying pressure and temperature, due to the viscous drag of the surrounding gas. For example, a piezoelectric bimorph cantilever was used to measure the pressure and temperature of the surrounding environment simultaneously, with an accuracy of 1 mbar and 0.03°C, respectively [87].

Cantilevers coated with a layer of a different thermal expansion coefficient bend when they are subjected to an external heat source. An aluminum-coated or gold-coated silicon cantilever can be used thus as a heat sensor, and micro-kelvin and picojoule sensitivities were achieved (Fig. 9.9c). These sensors allowed the investigation of exothermic chemical reactions, such as the conversion of hydrogen and oxygen to water vapor. Other potential applications include the study of reactions of photosensitive chemicals and the thermal analysis of small amounts of chemicals (hundreds of nanograms). With a heat sink deposited on the cantilever, the thermal properties of a material attached at the end of the cantilever can be studied by monitoring the deflection as a function of the temperature and comparing it to the data from a reference, similar to the concept of a differential scanning calorimeter (Fig. 9.9d) [88].

9.3.2 Nanolithography and High-Density Data Storage

Scanning probe-based nanolithography has long been explored as a way to increase the density of data storage systems, since it can write patterns nearly at the atomic scale. The first scanning probe-based data storage system used a silicon cantilever with an integrated tip heater to produce irreversible topographical features. The data were written by heating locally a polymer film with a hot tip while in contact and thus creating a small indent. Features as small as 40 nm in diameter with a pitch of 120 nm were obtained, yielding a potential density of 400 Gb/in.^{2}. The data were read with the same tip using the principle of thermal sensing. The tip is heated to a temperature that is below the minimum temperature required to make an indent and scanned over the polymer film. The thermal conductance between the cantilever and the polymer film depends on the distance between them. Over an indent, the cantilever–film distance decreases, improving the thermal conductance and therefore reducing the cantilever temperature and its resistance. Data are thus read by monitoring the cantilever resistance variations during the scan [89, 90]. The reading process can be improved by integrating piezoelectric sensors on the cantilever. When the tip is scanned across an indent, the topography leads to changes in the cantilever deflection while deforming the piezoelectric sensors placed under the cantilever. Stress variations on the piezoelectric sensors generate charges on their surface, which can be collected with electrodes (Fig. 9.10). This detection technique offers the advantages of lower power consumption and higher reading speed than thermal sensing [91].

Another scanning probe-based nanolithography technique is based on the local oxidation of a substrate, by forming a small electrochemical cell between the tip and the substrate immersed in an electrolyte. A write–read–erase data storage system was reported with the local oxidation of a tungsten oxide film. Local oxidation nanolithography usually yields larger features than other scanning probe-based techniques, but offers versatility in surface functionalization and the erasing capability required to design complex patterns [92].

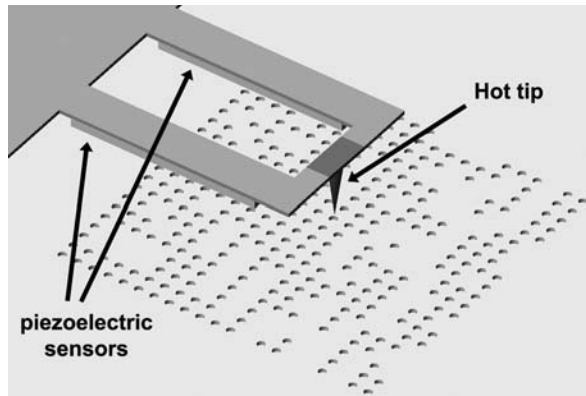


Fig. 9.10 Thermomechanical scanning probe-based data storage. The data are written on a polymer film with a hot tip in contact. Integrated piezoelectric sensors provide a feedback mechanism for the reading operation. (Reproduced with permission from [91]. Copyright 2007, Elsevier)

Scanning probes can also write nanodomains of inverted polarization in ferroelectric materials. The use of ferroelectric materials would significantly increase the data storage density, since the domain wall thickness is typically a few nanometers, much smaller than that in the ferromagnetic materials currently in use. The technique also offers other advantages such as non-volatility, a non-destructive reading process and rewritability. The nanodomains are created by applying a dc electric pulse between a sharp conducting tip and the substrate, thus creating an electric field perpendicular to the substrate which then induces a polarization switch of the domain under the tip parallel to the electric field. The data are then read with the same probe, by applying a voltage to the substrate and measuring the polarization-dependent piezoelectric response. Using this technique, arrays of nanodomains as small as 20 nm in diameter with a maximum density of 1.50 Tbit/in.^{2} were written on thin films of single-crystal lithium tantalite [93].

Currently, scanning probe-based nanolithography is limited by the slow writing and reading speed. In ideal conditions, AFMs operate at a microsecond time scale, while the magnetic data storage systems operate at a nanosecond time scale. The IBM Zurich research laboratory introduced the concept of parallel operation of scanning thermomechanical probes, by integrating 2D arrays of 32 $\times$ 32 cantilevers on a single chip ("Millipede" concept). The Millipede concept is based on a thermomechanical write/read process in a thin polymer film. The chip has two levels of wiring to form a multiplexed row/column addressing scheme. The rows are activated one by one, by supplying a heater current to all the cantilevers on a particular row. While a row is activated, data inputs (bits of "1" and "0") are delivered to the 32 columns. Only the cantilevers in the columns corresponding to "1" bits indent the polymer [90]. Three magnetic actuators control the distance between the entire chip and the polymer film (Fig. 9.11). Without individual tip–sample distance feedbacks, the flatness of the polymer film, the alignment of the cantilevers during the fabrication and the accurate leveling of the chip are critical. To alleviate this problem, a 128 $\times$ 128 silicon nitride probe array with integrated heaters and piezoelectric sensors on each cantilever was fabricated [91].

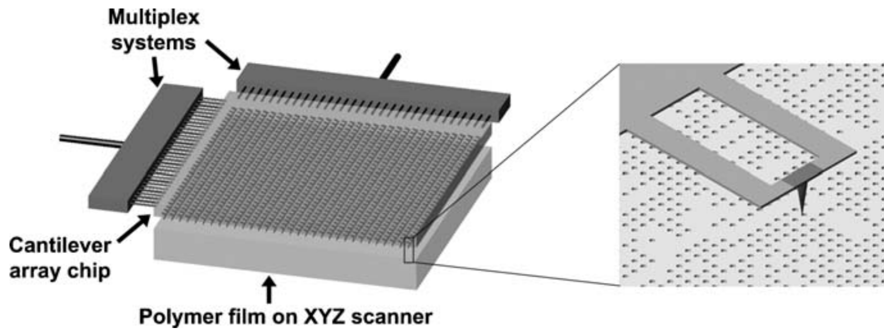


Fig. 9.11 Schematic of the "Millipede" concept. A 2D array of 32 × 32 cantilevers with integrated tip heaters writes data on a polymer film deposited on a XYZ scanner following a thermomechanical method. The cantilevers are addressed with multiplex systems. (Reproduced with permission from [89]. Copyright 1999, IEEE)

9.3.3 Optics and Telecommunications

A considerable effort has been made to develop and miniaturize high-frequency or optical device components for computing and wireless communications, where NEMS will play a major role. Optical connections and integrated circuits can significantly improve the performance of computers, since photons travel faster than electrons, and with less electromagnetic interference. Radio-frequency devices are already widely used in cell phones, cellular base station amplifiers and wireless local area networks and will soon be used in collision avoidance radars.

Radio-frequency MEMS act as switches or relays in the waveguides transmitting high-frequency signals, such as microwaves. Typical silicon-based MEMS have switching times in the microsecond range, which is too slow for high-speed applications. NEMSs have very low masses, so their switching times are expected to be in the nanosecond range. They can be integrated into coplanar waveguides, which consist of a central conductor placed between two semi-infinite grounded planes. Electromechanically activated nanotube tweezers are used as switches placed between the electrodes (Fig. 9.12). Without a dc voltage, the tweezers are open and there is no transmission along the waveguide. When a dc voltage is applied to the nanotubes, the nanotubes are in contact, thus creating a shortcut that allows the transmission of microwaves. The measured switching time in this device was 49 ns, three orders of magnitude lower than that of typical MEMS [94].

Tilting mirrors have been manufactured from silicon wafers using electron beam lithography. The moving part of the device is a $2 \times 2\ \mu m^2$ silicon wafer, suspended by 50 nm wide wires. The mirror is driven by resonant vibrations excited by an ac voltage between gold electrodes deposited on top of the moving

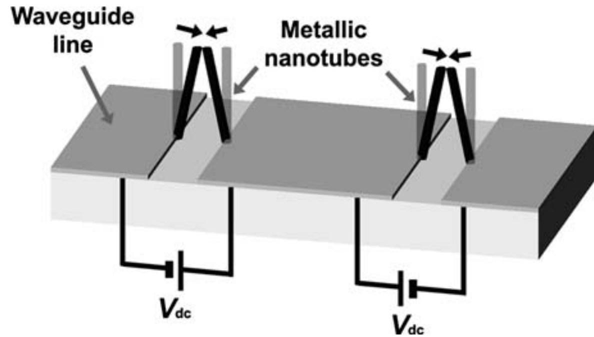


Fig. 9.12 Schematic of a coplanar waveguide activated by nanotube switches. When a voltage is applied between the metallic nanotubes, they come into contact, allowing the transmission of microwave signals. (Reused with permission from [94]. Copyright 2007, American Institute of Physics)

part and on the substrate [2]. The tilting mechanism may be coupled with a parallel-guiding mechanism that provides additional degrees of freedom along a linear or curved path (Fig. 9.13) [95].

Nanometer-range displacements are also useful for tuning the optical properties of photonic crystal structures, such as two parallel photonic crystal slabs (Fig. 9.14). Each slab is a high-index layer with a periodic array of air holes. The transmission and reflection coefficients of this structure are expected to vary significantly as a NEMS actuator modifies the gap between the slabs. Peaks in the transmission spectra shift to a higher or lower frequency, depending on the amplitude of the displacement [96].

9.3.4 Nanomanipulators

As device components are miniaturized to the nanoscale, the development of new manipulation and assembly tools becomes necessary. NEMSs are particularly desirable for the positioning, deformation and characterization of nanostructures. Nanotweezers were fabricated with carbon nanotubes attached to two independent electrodes deposited on a glass pipette. Applying voltages to

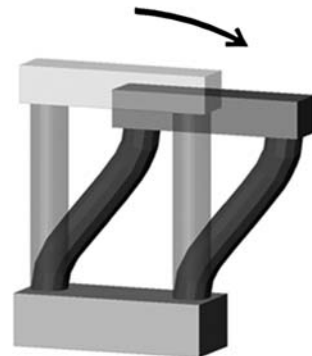


Fig. 9.13 Schematic of a parallel-guiding mechanism. A rigid coupler is guided by two carbon nanotubes deforming elastically. (Reused with permission from [95]. Copyright 2006, American Institute of Physics)

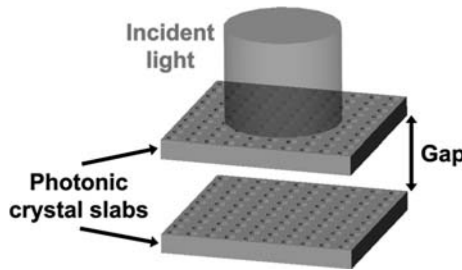


Fig. 9.14 Schematic of a gap-dependent photonic crystal structure. The transmission spectrum of the incident light through the device depends on the gap between the photonic crystal slabs. (Reused with permission from [96]. Copyright 2003, American Institute of Physics)

the electrodes closes or opens the nanotube arms. These nanotube actuators offer a reproducible elastic response and require lower actuating voltages (less than 10 V) than the previous systems made of silicon or tungsten. Nanotube nanotweezers were used to grasp individual polystyrene beads of about 500 nm in diameter, and also a GaAs nanowire to probe its electrical properties (Fig. 9.15) [97].

Aside from the small size of nanostructures, additional consideration should be given to delicate samples, especially in the biological field, such as cells, proteins or lipid bilayers. Soft materials, such as polymers, are more suited for biological applications, because of their mechanical flexibility, chemical versatility and low processing cost. The polymer-based NEMS can be operated in water, and polymers are suitable for photolithography and other scanning probe-based lithography techniques. Actuation of these NEMSs is controlled by electrochemical processes: the ion insertion (or removal) induces the expansion (or contraction) of the polymer film. Polypyrrole–gold bilayer actuators can potentially transport and isolate individual cells into microcavities, where their biological responses to specific proteins can be studied [98].

Other polymeric materials can be considered for additional functionality, including the thermosensitive polymer, poly(*N*-isopropylacrylamide) or PNIPAM, which undergoes reversible volume and wettability changes as the temperature varies. Red blood cells were stretched or compressed in a PNIPAM gel

Fig. 9.15 Schematic of nanotube nanotweezers. Two carbon nanotubes are attached to two independent electrodes deposited on a glass micropipette. The nanotweezers are closed by applying a voltage between the electrodes. (From [97]. Reproduced with permission from AAAS)

cavity as the gel volume varied with temperature [99]. Since the deformation of cells affects their biological response, these thermosensitive polymers can be integrated in NEMS to act as electrical switches for their adsorption or their biological function.

9.3.5 Catalysis

Molecular dynamics studies of carbon nanotubes showed that all mechanical deformation modes, including axial compression/tension, torsion and bending, significantly affect the binding of atoms and radicals [100]. Structural variations of catalysts are also expected to affect the chemical reaction rates and potentially the structures of the reaction products, notably their chirality. Theoretical studies suggested the possibility to tune the activity of catalysts by attaching them to a surface that can be deformed reversibly. As an example, the configuration variations and the catalytic activity of a chiral molecule adsorbed on the sidewall of a carbon nanotube were studied as the nanotube was twisted. Carbon nanotubes are known to be extremely resilient and can sustain large elastic deformations. A small twist of the nanotube affects the binding energy of the catalyst and its axis by tens of degrees, which then prevents the formation of a specific configuration of the product [101].

Recently, a torsional pendulum based on an individual carbon nanotube was fabricated by electron beam lithography. The carbon nanotube acts as a torsional spring and support for the moving part. The application of an electric field rotates the moving part, which then induces the elastic torsion of the nanotube (Fig. 9.16). The nanotube remains intact after the electric field is turned off, even when the moving part is rotated by 180° [102]. Such a system can be used as an electrically switchable catalytic site or for chiral recognition, since carbon nanotubes can be chemically functionalized and they also present a high surface-to-volume ratio.

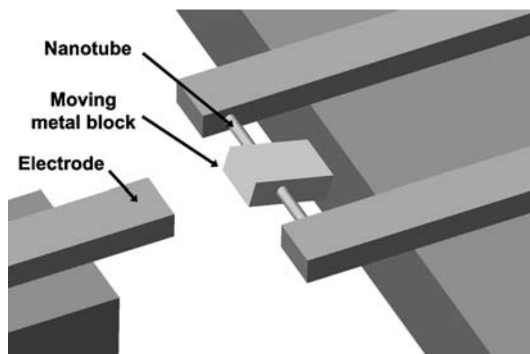


Fig. 9.16 A moving metal block is suspended by an individual carbon nanotube. The moving block is turned by applying a voltage between the electrode and the nanotube support. (From [102]. Reproduced with permission from AAAS)

9.3.6 Electrical Power Generation

Advances in miniaturization have led to considerable reduction in power consumption in nanodevices. However, most NEMSs still require an external power source, which can restrict their applications, notably in the biomedical field where non-invasive techniques are particularly desirable. Exploiting motions or mechanical vibrations to generate electrical power is an attractive prospect, considering its low cost. Different types of systems have been developed to generate electrical power from human body motion. One system is based on an eccentric rotor: a human body motion changes the position of the rotor around its axis of rotation or makes it swing. When the body motion is oscillatory, as in a walk or a run, the self-excited rotation of the rotor can be used to generate electrical power. A different system exploits the resonant vibrations of a magnet placed inside a coil and suspended by springs. The power output reaches a maximum when the frequency of the oscillatory motion matches the resonance frequency of the system, which can be adjusted by tuning the mass and the elastic constant of the springs [103].

A power generator based on an array of piezoelectric ZnO nanowires was also developed to convert mechanical vibrations induced by an ultrasonic wave into electricity (Fig. 9.17). An array of vertically aligned ZnO nanowires is grown on a GaN substrate covered with a ZnO film, which serves as the bottom electrode, and then covered by a silicon wafer with triangular trenches coated with a layer of platinum, acting as the top electrode. The ultrasonic wave drives the top electrode up and down, with the triangular trenches inducing a lateral deflection of the ZnO nanowires. The deflection leads to a difference in piezoelectric potential between the stretched side (positive potential) and compressed

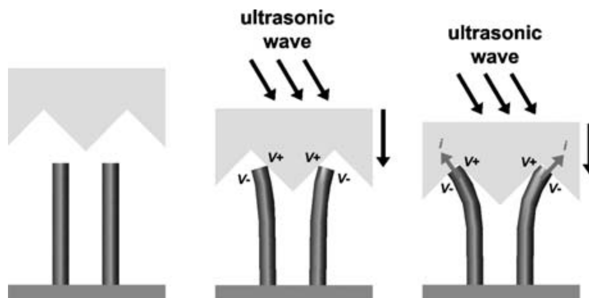


Fig. 9.17 Electrical power generation driven by ultrasonic waves. (**a**) The nanogenerator is based on an array of vertical ZnO nanowires. The top electrode has triangular trenches and is covered by a thin Pt film. (**b**) An ultrasonic wave drives the top electrode down, bending the ZnO nanowires and creating opposite piezoelectric potentials on the stretched and compressed sides of the nanowires. (**c**) The top electrode is driven further down, to the point where the compressed side of the nanowires is also in contact with the top electrode, resulting in a piezoelectric discharge and electrical current flow. (From [104]. Reproduced with permission from AAAS)

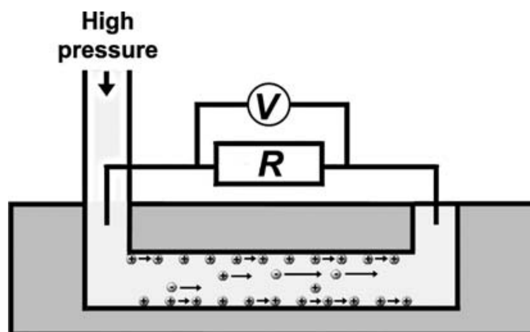


Fig. 9.18 Electrical power generated by the transport of ions in a nanofluidic channel. A high pressure drives the fluid flow, carrying counterions near the channel walls, and thus generating an electrical current. (Reproduced with permission from [105]. Copyright 2007, American Chemical Society)

side (negative potential) of the nanowires. Charges are then accumulated at the interface between the top electrode and the stretched side of the nanowires. Upon further reduction of the gap between the electrodes, the compressed side of the nanowire with a negative potential makes contact with the top electrode, resulting in the release of the accumulated charges and an electrical current. This small generator can potentially be interfaced with implantable biodetectors [104].

Finally, another promising way to generate electrical power is to exploit the flow of electrical charges in a nanofluidic channel. A pressure-induced fluid flow carries the counter charges that are accumulated in the double layer near the channel walls, generating an electrical current along the flow (Fig. 9.18). The energy conversion efficiency was found to depend on the ion concentration and the size of the nanochannel [105]. Using this concept, a nanodevice could be implanted and use the blood stream to power biodetectors or drug delivery systems.

9.4 Summary and Outlook

The development of new techniques, mainly based on AFM and TEM, has enabled the characterization of mechanical properties of nanostructures and revealed the importance of surface effects and defects on their size dependence. Macroscopic phenomena, such as the motion of dislocations, can still be observed and describe the mechanical behaviors of nanostructures with dimensions below 100 nm. At a smaller scale, the influence of surface stresses and defects becomes predominant and they can lead to the stiffening or softening of nanostructures. Also, phase transitions are predicted to occur during the plastic deformation of nanometer-sized single crystals, but so far they have not been observed experimentally. A detailed study of the mechanical properties and their influence on the other physical/chemical properties is essential for the application of NEMS in biological/chemical sensing, data storage, telecommunications and electrical power generation.

Before the massive production of new devices, three main issues must be resolved to tailor the mechanical properties of nanostructures for their desired application: the control over their morphology and structure during the synthesis, the development of new methods to manipulate and position them and the combination of different techniques for the complete characterization of their structure–properties relationship. For example, the use of catalysts and templates is intensively explored for the mass production of single-wall carbon nanotubes with a specific chirality [106]. New scanning probe-based nanolithography techniques are developed to create patterns with surface chemistry to immobilize, position and align nanostructures, particularly in biology [107]. Finally, the combination of X-ray microscopy [108] with spectroscopy would allow a local 3D chemical and structural analysis of soft nanostructures under deformation.

Acknowledgments The authors acknowledge the financial support from the DoE (grant no. DE-FG02-06ER46293) and NSF (grant no. DMR-0120967 and no. DMR-0405319).

References

1. Kuchibhatla SVNT, Karakoti AS, Bera D, Seal S (2007) One dimensional nanostructured materials. Prog Mater Sci 52:699–913
2. Craighead HG (2000) Nanoelectromechanical systems. Science 290:1532–1535
3. Schwab KC, Roukes ML (2005) Putting mechanics into quantum mechanics. Phys Today 58:36–42
4. Tombler TW, Zhou C, Alexseyev L, Kong J, Dai H, Liu L, Jayanthi CS, Tang M, Wu S (2000) Reversible electromechanical characteristics of carbon nanotubes under local-probe manipulation. Nature 405:769–772
5. Cao J, Wang Q, Dai H (2003) Electromechanical properties of metallic, quasimetallic, and semiconducting carbon nanotubes under stretching. Phys Rev Lett 90:157601 1–4
6. Yamamoto T, Watanabe K (2006) Nonequilibrium Green's function approach to phonon transport in defective carbon nanotubes. Phys Rev Lett 96:255503 1–4
7. Delimitis A, Komninou P, Dimitrakopulos GP, Kehagias T, Kioseoglou J, Karakostas T, Nouet G (2007) Strain distribution of thin InN epilayers grown on (0001) GaN templates by molecular beam epitaxy. Appl Phys Lett 90:061920 1–3
8. Roder C, Einfeldt S, Figge S, Paskova T, Hommel D, Paskov PP, Monemar B, Behn U, Haskell BA, Fini PT, Nakamura S (2006) Stress and wafer bending of a-plane GaN layers on r-plane sapphire substrates. J Appl Phys 100:103511 1–11
9. Seravalli L, Minelli M, Frigeri P, Franchi S, Guizzetti G, Patrini M, Ciabattoni T, Geddo M (2007) Quantum dot strain engineering of InAs/InGaAs nanostructures. J Appl Phys 101:024313 1–8
10. Lucas M, Young RJ (2004) Effect of uniaxial strain deformation upon the Raman radial breathing modes of single-wall carbon nanotubes in composites. Phys Rev B 69:085405 1–9
11. Sandler J, Shaffer MSP, Windle AH, Halsall MP, Montes-Morn MA, Cooper CA, Young RJ (2003) Variations in the Raman peak shift as a function of hydrostatic pressure for various carbon nanostructures: a simple geometric effect. Phys Rev B 67:035417 1–8
12. Bettinger HF (2005) The reactivity of defects at the sidewalls of single-walled carbon nanotubes: the Stone-Wales defect. J Phys Chem B 109:6922–6924

13. Robinson JA, Snow ES, Badescu, Reinecke TL, Perkins FK (2006) Role of defects in single-walled carbon nanotube chemical sensors. Nano Lett 6:1747–1751
14. Li TD, Gao J, Szoszkiewicz R, Landman U, Riedo E (2007) Structured and viscous water in subnanometer gaps. Phys Rev B 75:115415 1–6
15. Butt HJ, Cappella B, Kappl M (2005) Force measurements with the atomic force microscope: Technique, interpretation and applications. Surf Sci Rep 59:1–152
16. Poncharal P, Wang ZL, Ugarte D, de Heer WA (1999) Electrostatic deflections and electromechanical resonances of carbon nanotubes. Science 283:1513–1516
17. Liu KH, Wang WL, Xu Z, Liao L, Bai XD, Wang EG (2006) In situ probing mechanical properties of individual tungsten oxide nanowires directly grown on tungsten tips inside transmission electron microscope. Appl. Phys Lett 89:221908 1–3
18. Nam CY, Jaroenapibal P, Tham D, Luzzi DE, Evoy S, Fischer JE (2006) Diameter-dependent electromechanical properties of GaN nanowires. Nano Lett 6:153–158
19. Miyake K, Satomi N, Sasaki S (2006) Elastic modulus of polystyrene film from near surface to bulk measured by nanoindentation using atomic force microscopy. Appl Phys Lett 89:031925 1–3
20. Shin MK, Kim SI, Kim SJ, Kim SK, Lee H, Spinks GM (2006) Size-dependent elastic modulus of single electroactive polymer nanofibers. Appl Phys Lett 89:231929 1–3
21. Vigolo B, Poulin P, Lucas M, Launois P, Bernier P (2002) Improved structure and properties of single-wall carbon nanotube spun fibers. Appl Phys Lett 81:1210–1212
22. Wu B, Heidelberg A, Boland JJ (2005) Mechanical properties of ultrahigh-strength gold nanowires. Nat Mater 4:525–529
23. Haque MA, Saif MTA (2004) Deformation mechanisms in free-standing nanoscale thin films: a quantitative in situ transmission electron microscope study. Proc Natl Acad Sci USA 101:6335–6340
24. Villain P, Goudeau P, Renautl PO, Badawi KF (2002) Size effect on intragranular elastic constants in thin tungsten films. Appl Phys Lett 81:4365–4367
25. Palaci I, Fedrigo S, Brune H, Klinke C, Chen M, Riedo E (2005) Radial elasticity of multiwalled carbon nanotubes. Phys Rev Lett 94:175502 1–4
26. Yao N, Lordi V (1998) Young's modulus of single-walled carbon nanotubes. J Appl Phys 84:1939–1943
27. Huang JY, Chen S, Wang ZQ, Kempa K, Wang YM, Jo SH, Chen G, Dresselhaus MS, Ren ZF (2006) Superplastic carbon nanotubes. Nature 439:281
28. Yakobson BI, Brabec CJ, Bernholc J (1996) Nanomechanics of carbon tubes: instabilities beyond linear regime. Phys Rev Lett 76:2511–2514
29. Champion Y, Langlois C, Guérin-Mailly S, Langlois P, Bonnentien JL, Hÿtch MJ (2003) Near-perfect elastoplasticity in pure nanocrystalline copper. Science 300:310–311
30. Lu L, Sui ML, Lu K (2000) Superplastic extensibility of nanocrystalline copper at room temperature. Science 287:1463–1466
31. Kulkarni AJ, Zhou M, Sarasamak K, Limpijumnong S (2006) Novel phase transformation in ZnO nanowires under tensile loading. Phys Rev Lett 97:105502 1–4
32. Zhou LG, Huang H (2004) Are surfaces elastically softer or stiffer? Appl Phys Lett 84:1940–1942
33. Liang H, Upmanyu M, Huang H (2005) Size-dependent elasticity of nanowires: nonlinear effects. Phys Rev B 71:241403(R) 1–4
34. Kong XY, Wang ZL (2003) Spontaneous polarization-induced nanohelixes, nanosprings, and nanorings of piezoelectric nanobelts. Nano Lett 3:1625–1634
35. Sun CT, Zhang H (2003) Size-dependent elastic moduli of platelike nanomaterials. J Appl Phys 93:1212–1218
36. Miller RE, Shenoy VB (2000) Size-dependent elastic properties of nanosized structural elements. Nanotechnology 11:139–147
37. Ji C, Park HS (2006) Geometric effects on the inelastic deformation of metal nanowires. Appl Phys Lett 89:181916 1–3

38. Park HS, Gall K, Zimmerman JA (2005) Shape memory and pseudoelasticity in metal nanowires. Phys Rev Lett 95:255504 1–4
39. Renautl PO, Le Bourhis E, Villain P, Goudeau P, Badawi KF, Faurie D (2003) Measurement of the elastic constants of textured anisotropic thin films from x-ray diffraction data. Appl Phys Lett 83:473–475
40. Cuenot S, Frétigny C, Demoustier-Champagne S, Nysten B (2004) Surface tension effect on the mechanical properties of nanomaterials measured by atomic force microscopy. Phys Rev B 69:165410 1–5
41. Chen CQ, Shi Y, Zhang YS, Zhu J, Yan YJ (2006) Size dependence of Young's modulus in ZnO nanowires. Phys Rev Lett 96:075505 1–4
42. Jing GY, Duan HL, Sun XM, Zhang ZS, Xu J, Li YD, Wang JX, Yu DP (2006) Surface effects on elastic properties of silver nanowires: contact atomic-force microscopy. Phys Rev B 73:235409 1–6
43. Lu L, Shen Y, Chen X, Qian L, Lu K (2004) Ultrahigh strength and high electrical conductivity in copper. Science 304:422–426
44. Arzt E (1998) Size effects in materials due to microstructural and dimensional constraints: a comparable review. Acta Mater 46:5611–5626
45. Uchic MD, Dimiduk DM, Florando JN, Nix WD (2004) Sample dimensions influence strength and crystal plasticity. Science 305:986–989
46. Kiely JD, Hwang RQ, Houston JE (1998) Effect of surface steps on the plastic threshold in nanoindentation. Phys Rev Lett 81:4424–4427
47. Enomoto K, Kitakata S, Yasuhara T, Ohtake N, Kuzumaki T, Mitsuda Y (2006) Measurement of Young's modulus of carbon nanotubes by nanoprobe manipulation in a transmission electron microscope. Appl Phys Lett 88:153115 1–3
48. Mukhopadhyay AK, Chaudhuri MR, Seal A, Dalui SK, Banerjee M, Phani KK (2001) Mechanical characterization of microwave sintered zinc oxide. Bull Mater Sci 24:125–128
49. Cuenot S, Demoustier-Champagne S, Nysten B (2000) Elastic modulus of polypyrrole nanotubes. Phys Rev Lett 85:1690–1693
50. Lucas M, Mai W, Yang R, Wang ZL, Riedo E (2007) Aspect ratio dependence of the elastic properties of ZnO nanobelts. Nano Lett 7:1314–1317
51. Kucheyev, Bradby JE, Williams JS, Jagadish C, Swain MV (2001) Mechanical deformation of single-crystal ZnO. Appl Phys Lett 80:956–958
52. Bradby JE, Kucheyev SO, Williams JS, Jagadish C, Swain MV, Munroe P, Phillips MR (2002) Contact-induced defect propagation in ZnO. Appl Phys Lett 80:4537–4539
53. Corcoran SG, Colton RJ, Lilleodden ET, Gerberich WW (1997) Anomalous plastic deformation at surfaces: nanoindentation of gold single crystals. Phys Rev B 55:16057–16060
54. Richter A, Wolf B, Nowicki M, Smith R, Usov IO, Valdez JA, Sickafus K (2006) Multicycling nanoindentation in MgO single crystals before and after ion irradiation. J Phys D: Appl Phys 39:3342–3349
55. Kis A, Csányi G, Salvetat JP, Lee T, Couteau E, Kulik AJ, Benoit W, Brugger J, Forr (2004) Reinforcement of single-walled carbon nanotube bundles by intertube bridging. Nat Mater 3:153–157
56. Leach AM, McDowell M, Gall K (2007) Deformation of top-down and bottom-up silver nanowires. Adv Funct Mater 17:43–53
57. Wu B, Heidelberg A, Boland JJ, Sader JE, Sun X, Li Y (2006) Microstructure-hardened silver nanowires. Nano Lett 6:468–472
58. Fleck NA, Hutchinson JW (1993) A phenomenological theory for strain gradient effects in plasticity. J Mech Phys Solids 41:1825–1857
59. Nix WD, Gao H (1998) Indentation size effects in crystalline materials: a law for strain gradient plasticity. J Mech Phys Solids 46:411–425
60. Kiely JD, Houston JE (1998) Nanomechanical properties of Au(111), (001), and (110) surfaces. Phys Rev B 57:12588–12594

61. Kulkarni AJ, Zhou M, Ke FJ (2005) Orientation and size dependence of the elastic properties of zinc oxide nanobelts. Nanotechnology 16:2749–2756
62. Trabelsi S, Albouy PA, Rault J (2002) Stress-induced crystallization around a crack tip in natural rubber. Macromolecules 35:10054–10061
63. Diao J, Gall K, Dunn M (2003) Surface-stress-induced phase transformation in metal nanowires. Nat Mater 2:656–660
64. Zhang L, Huang H (2006) Young's moduli of ZnO nanoplates: *ab initio* determinations. Appl Phys Lett 89:183111 1–3
65. Wong EW, Sheehan PE, Lieber CM (1997) Nanobeam mechanics: elasticity, strength, and toughness of nanorods and nanotubes. Science 277:1971–1975
66. Salvetat JP, Briggs GAD, Bonard JM, Bacsa RR, Kulik AJ, Stckli T, Burnham NA, ForrÓ L (1999) Elastic and shear moduli of single-walled carbon nanotube ropes. Phys Rev Lett 82:944–947
67. Shanmugham S, Jeong J, Alkhateeb A, Aston DE (2005) Polymer nanowire elastic moduli measured with digital pulsed force mode AFM. Langmuir 21:10214–10218
68. Chen Y, Dorgan BL Jr, McIlroy DN, Aston DE (2006) On the importance of boundary conditions on nanomechanical bending behavior and elastic modulus determination of silver nanowires. J Appl Phys 100:104301 1–7
69. San Paulo A, Bokor J, Howe RT, He R, Yang P, Gao D, Carraro C, Maboudian R (2005) Mechanical elasticity of single and double clamped silicon nanobeams fabricated by the vapor-liquid-solid method. Appl Phys Lett 87:053111 1–3
70. Mai W, Wang ZL (2006) Quantifying the elastic deformation behavior of bridged nanobelts. Appl Phys Lett 89:073112 1–3
71. Lulevich V, Zink T, Chen HY, Liu FT, Liu G (2006) Cell mechanics using atomic force microscopy-based single-cell compression. Langmuir 22:8151–8155
72. Li X, Chen W, Zhan Q, Dai L, Sowards L, Pender M, Naik RR (2006) Direct measurements of interactions between polypeptides and carbon nanotubes. J Phys Chem B 110:12621–12625
73. Song J, Wang X, Riedo E, Wang ZL (2005) Elastic property of vertically aligned nanowires. Nano Lett 5:1954–1958
74. Lucas M, Mai WJ, Yang RS, Wang ZL, Riedo E (2007) Size dependence of the mechanical properties of ZnO nanobelts. Philos Mag 87:2135–2141
75. Yamanaka K (1996) UFM observation of lattice defects in highly oriented pyrolytic graphite. Thin Solid Films 273:116–121
76. Rabe U, Arnold W (1994) Acoustic microscopy by atomic force microscopy. Appl Phys Lett 64:1493–1495
77. Zheng Y, Geer RE, Dovidenko K, Kopycinska-Müller M, Hurley DC (2006) Quantitative nanoscale modulus measurements and elastic imaging of SnO_2 nanobelts. J Appl Phys 100:124308 1–6
78. Treacy MMJ, Ebbesen TW, Gibson JM (1996) Exceptionally high Young's modulus observed for individual carbon nanotubes. Nature 381:678–680
79. Gaillard J, Skove M, Rao AM (2005) Mechanical properties of chemical vapor deposition-grown multiwalled carbon nanotubes. Appl Phys Lett 86:233109 1–3
80. Yu MF, Lourie O, Dyer MJ, Moloni K, Kelly TF, Ruoff RS (2000) Strength and breaking mechanism of multiwalled carbon nanotubes under tensile load. Science 287:637–640
81. Ding W, Calabri L, Chen X, Kohlhaas KM, Ruoff RS (2006) Mechanics of crystalline boron nanowires. Comp Sci Tech 66:1112–1124
82. Ashkin A, Dziedzic JM, Bjorkholm JE, Chu S (1986) Observation of a single-beam gradient force optical trap for dielectric particles. Opt Lett 11:288–290
83. Bustamante C, Bryant Z, Smith S (2003) Ten years of tension: single-molecule DNA mechanics. Nature 421:423–427

84. Li Y, Lim HS, Ng SC, Wang ZK, Kuok MH, Vekris E, Kitaev V, Peiris FC, Ogin GA (2006) Micro-Brillouin scattering from a single isolated nanosphere. Appl Phys Lett 88:023112 1–3
85. Wu G, Datar RH, Hansen KM, Thundat T, Cote RJ, Majumdar A (2001) Bioassay of prostate-specific antigen (PSA) using microcantilevers. Nat Biotech 19:856–860
86. Li M, Tang HX, Roukes ML (2007) Ultra-sensitive NEMS-based cantilevers for sensing, scanned probe and very high-frequency applications. Nat Nanotech 2:114–120
87. Mortet V, Petersen R, Haenen K, D'Olieslaeger M (2006) Wide range pressure sensor based on a piezoelectric bimorph microcantilever. Appl Phys Lett 88:133511 1–3
88. Lang HP, Hegner M, Meyer E, Gerber C (2002) Nanomechanics from atomic resolution to molecular recognition based on atomic force microscopy technology. Nanotechnology 13:R29–R36
89. Despont M, Brugger J, Drechsler U, Dürig U, Häberle W, Lutwyche M, Rothuizen H, Stutz R, Widmer R, Rohrer H, Binnig G, Vettiger P (1999) VLSI-NEMS chip for AFM data storage. Technical Digest, 12th IEEE International Micro Electro Mechanical Systems Conference (MEMS'99), Orlando, FL, January 1999, pp. 564–569
90. Vettiger P, Despont M, Dreschler U, Durig U, Haberle W, Lutwyche MI, Rothuizen HE, Stutz R, Widmer R, Binnig GK (2000) The Millipede – More than one thousand tips for future AFM data storage. IBM J Res Dev 44:323–340
91. Nam HJ, Kim YS, Lee CS, Jin WH, Jang SS, Choa IJ, Bua JU, Choi WB, Choi SW (2007) Silicon nitride cantilever array integrated with silicon heaters and piezoelectric detectors for probe-based data storage. Sens Actuators A 134:329–333
92. Turyan I, Krasovec UO, Orel B, Saraidorov T, Reisfeld R, Mandler D (2000) "Writing-reading-erasing" on tungsten oxide films using the scanning electrochemical microscope. Adv Mater 12:330–333
93. Cho Y, Fujimoto K, Hiranaga Y, Wagatsuma Y, Onoe A, Terabe K, Kitamura K (2002) Tbit/inch2 ferroelectric data storage based on scanning nonlinear dielectric microscopy. Appl Phys Lett 81:4401–4403
94. Dragoman M, Takacs A, Muller AA, Hartnagel H, Plana R, Grenier K, Dubuc D (2007) Nanoelectromechanical switches based on carbon nanotubes for microwave and millimeter waves. Appl Phys Lett 90:113102 1–3
95. Culpepper ML, DiBiasio CM, Panas RM, Magleby S, Howell LL (2006) Simulation of a carbon nanotube-based compliant parallel-guiding mechanism: A nanomechanical building block. Appl Phys Lett 89:203111 1–3
96. Suh W, Yanik MF, Solgaard O, Fan S (2003) Displacement-sensitive photonic crystal structures based on guided resonance in photonic crystal slabs. Appl Phys Lett 82:1999–2001
97. Kim P, Lieber CM (1999) Nanotube nanotweezers. Science 286:2148–2150
98. Jager EWH, Smela E, Ingans O (2000) Microfabricating conjugated polymer actuators. Science 290:1540–1545
99. Pelah A, Seemann R, Jovin TM (2007) Reversible cell deformation by a polymeric actuator. J Am Chem Soc 129:468–469
100. Mylvaganam K, Zhang LC (2006) Deformation-promoted reactivity of single-walled carbon nanotubes. Nanotechnology 17:410–414
101. Wang B, Král P, Thanopulos I (2006) Docking of chiral molecules on twisted and helical nanotubes: nanomechanical control of catalysis. Nano Lett 6:1918–1921
102. Meyer JC, Paillet M, Roth S (2005) Single-molecule torsional pendulum. Science 309:1539–1541
103. Sasaki K, Osaki Y, Okazaki J, Hosaka H, Itao K (2005) Vibration-based automatic power-generation system. Microsyst Technol 11:965–969
104. Wang X, Song J, Liu J, Wang ZL (2007) Direct-current nanogenerator driven by ultrasonic waves. Science 316:102–105

105. Van der Heyden FHJ, Bonthuis DJ, Stein D, Meyer C, Dekker C (2007) Power generation by pressure-driven transport of ions in nanofluidic channels. Nano Lett 7:1022–1025
106. Smalley RE, Li Y, Moore VC, Price BK, Colorado R, Schmidt HK, Hauge RH, Barron AR, Tour JM (2006) Single wall carbon nanotube amplification: en route to a type-specific growth mechanism. J Am Chem Soc 128:15824–15829
107. Szoszkiewicz R, Okada T, Jones SC, Li TD, King WP, Marder SR, Riedo E (2007) High-speed, sub-15 nm feature size thermochemical nanolithography. Nano Lett 7:1064–1069
108. Pfeifer MA, Williams GJ, Vartanyants IA, Harder R, Robinson IK (2006) Three-dimensional mapping of a deformation field inside a nanocrystal. Nature 442:63–66

Chapter 10
Classical and Quantum Optics of Semiconductor Nanostructures

Walter Hoyer, Mackillo Kira, and Stephan W. Koch

10.1 Introduction

Optical properties of semiconductor nanostructures are widely studied both experimentally and theoretically. They are interesting from an application point of view while they also provide an ideal playground to study Coulomb effects, light–matter interaction, and so forth. For the theoretical modeling, the strongly interacting charge carriers inside a semiconductor present a considerable challenge. This is intensified if also the electromagnetic radiation and potentially also the lattice vibrations have to be treated quantum mechanically. Direct solutions of, e.g., the Schrödinger equation are completely out of question, and a successful theoretical approach has to find consistent methods of truncating the infinite hierarchy problem caused by the interaction. In particular, Coulomb correlations have to be dealt with on the same footing as phonon or photon correlations.

Our theoretical approach is based on the Heisenberg equation of motion where the precise density matrix of the total system never has to be known. Instead, we will show in this article how quantum mechanically correct equations of motion can be derived for any quantities of interest as soon as the total system Hamiltonian is known. Thus, the precise knowledge of the Hamilton operator is of utmost importance and it should therefore include all relevant interaction mechanisms of all interacting quasi-particles of interest. Due to this prominent role of the Hamiltonian, we have split this article into two parts. The first two sections deal exclusively with the derivation of the semiconductor Hamiltonian of a nanostructure interacting with both a quantized light field and quantized lattice vibrations. While Section 10.2 deals with the contributions of the non-interacting quasi-particles and introduces important concept of the electronic band structure, the interaction contributions are discussed in Section 10.3. In Section 10.4, we calculate the elementary Heisenberg equation

S.W. Koch
Department of Physics and Material Sciences Center, Philipps-University Marburg, Renthof 5, D-35032 Marburg, Germany
e-mail: stephan.w.koch@physik.uni-marburg.de

A. Korkin, F. Rosei (eds.), *Nanoelectronics and Photonics*,
DOI: 10.1007/978-0-387-76499-3_10, © Springer Science+Business Media, LLC 2008

of motion for electronic, photon, and phonon operator and introduce the concept of the cluster expansion. Up to this point, the article is very explicit and tutorial and will give the reader a thorough introduction and understanding of the underlying concepts of our method. Having worked through this first tutorial part, the reader should be able to go ahead and calculate relevant subsystems of equations all by himself.

In the second half of the article, we present a few exemplary applications of the theory. Large parts of the discussions and figures in this second part are taken from Kira and Koch [1] where more details on the more advanced topics can be found. Our examples are divided into three bigger blocks. In Section 10.5, we begin by examining the typical absorption spectrum of a semiconductor heterostructure. We introduce the concept of coherent excitons as the relevant electronic quasi-particle at low carrier densities, and the generalization of the excitonic concept as well as the inclusion of microscopic carrier scattering for elevated densities. Such a description is valid for typical pump-probe setups. They can be described and understood by a semiclassical treatment with a classical electromagnetic field. In Section 10.6, we then turn to the more quantum-optical effects. Examples in the present article are photoluminescence spectra after non-resonant excitation as well as some quantum correlations of photons emitted from a quantum well into two different directions. This latter example is somewhat analogous to the traditional which-way experiment of quantum mechanics. While the incoherent spectra in Section 10.6 are always dominated by a strong excitonic resonance, it is well known that such an excitonic peak does not provide an unambiguous signature for true incoherent excitons. In Section 10.7, we therefore discuss in more detail how true exciton populations can be detected. We will show that the true analogue to probing atoms with optical fields is given by THz absorption in semiconductors because THz radiation lies in the proper electromagnetic frequency range to induce transitions between different excitonic levels.

10.2 Quantization of Quasi-particles in Semiconductors

The key ingredient to every quantum theory is the Hamilton operator $\hat{H}$ which defines the eigenstates of any quantum system, and which also determines the dynamics of all relevant quantities of interest. In particular in the Heisenberg picture, the Hamilton operator is needed in order to calculate the Heisenberg equation of motion. Thus, as a first step in our study of semiclassical and quantum-optical effects in semiconductors, we have to make a careful decision as to which phenomena we want to include into $\hat{H}$ such that we obtain a consistent and at the same time a technically feasible treatment.

Clearly, any discussion of quantum-optical features needs at least a quantum-mechanical description of the light. For the active semiconductor material, we start from a level where the atomic constituents that make up the solids are

periodically arranged in a lattice. As usual in solid-state physics, we assume that each atom can be subdivided into the weakly bound "outer" electrons and the "inner" core electrons that are strongly bound to the specific nucleus. Since nucleus and core electrons are usually not involved in the optical processes under discussion in this article, we will consider them as single ionic entities. The optical and electronic transitions only involve the outermost electrons, which we also refer to as (charge) *carriers*. Under the influence of applied fields, these carriers can perform transitions from one electronic state to another, i.e., they are optically active. Thus, we also need a quantum-mechanical description of carriers in order to describe quantum-optical effects originating from the interaction between the active electrons and the light.

In general, the dynamics of the active electrons is much faster than the ionic motion. As a result, these electrons rapidly adjust themselves to the momentary configuration of the ionic crystal such that one can treat the lattice dynamics independently of the electronic subsystem. This approach is commonly known as the Born–Oppenheimer approximation [2]. In order to understand the basic consequences of the periodic lattice, one actually does not even need to describe the ions individually on a microscopic level; one can rather adopt a mean-field approach, where the ions in the crystal lattice provide a periodic mean-field potential for the electrons. The electronic band structure resulting from the lattice periodic potential is discussed in Section 10.2.3.

Besides the coupling between light and carriers, the semiconductor excitations are subject also to other important interaction mechanisms. Since the carriers are charged particles, the unavoidable Coulomb force couples a single carrier to all other carriers due to the long-range nature of the interaction. This makes a semiconductor a genuinely Coulomb correlated many-body system which has to be analyzed properly in order to understand its optical and transport properties.

The active electrons can also interact with vibrational states, i.e., phonon excitations of the lattice resulting from the oscillatory distortions of the periodic crystal structure. As a result, the electronic system is directly coupled to the ionic environment. Sometimes, it is justified to describe these lattice vibrations as a reservoir (heat bath) with a well-defined temperature. The detailed analysis of this interaction is based on the microscopic description of the ionic motion in the lattice; this will be discussed in Sections 10.2.8 and 10.3.4.

The active electrons can also interact with the disorder-generated irregularities in the lattice. However, semiconductor manufacturing technologies have advanced tremendously during the past few decades. As a matter of fact, many state-of-the-art structures have reached a quality where one can observe effects much beyond the limitations of disorder. Since investigations with this kind of samples are most attractive from the point of view of the quantum optics, we mainly focus on semiconductor systems without disorder.

These considerations set a clear guideline on how the system Hamiltonian has to be constructed. Since the general formulation is most transparent in the first quantization, we start at this level in order to introduce the basic concepts.

Once the general properties are known, we use the formalism of the second quantization to obtain a description that is more suitable for the analysis of the complicated many-body problem.

10.2.1 System Hamiltonian in First Quantization

In the first step of our analysis, we assume that N optically active electrons are moving in a periodic potential provided by the positively charged ions that are rigidly arranged in the perfect crystal structure of the solid. The carriers are also coupled to each other via the Coulomb interaction. The optical transitions follow from the interaction of light with the carriers, which is described by the general minimal-substitution Hamiltonian [3]. If a transversal electromagnetic field interacts with N charged carriers, the system Hamiltonian has the general form

$$\hat{H}_N = \sum_{j=1}^{N} \left\{ \frac{1}{2m_0} [\hat{\mathbf{p}}_j - Q\hat{\mathbf{A}}(\mathbf{r}_j)]^2 + V_{\mathrm{L}}(\mathbf{r}_j) \right\} + \frac{1}{2} \sum_{i \neq j}^{N} V(|\mathbf{r}_i - \mathbf{r}_j|) + \hat{H}_{\mathrm{em}} + \hat{H}_{\mathrm{ph,}} \tag{10.1}$$

where $\hat{\mathbf{p}}_j$ and $\hat{\mathbf{r}}_j$ are, respectively, the canonical momentum and position operators of an electron j with charge Q and free-electron mass m_0. The lattice periodic potential is denoted as $V_{\mathrm{L}}(\mathbf{r})$ and

$$V(\mathbf{r}) = \frac{|e|^2}{4\pi\varepsilon_0} \frac{1}{\varepsilon|\mathbf{r}|} \tag{10.2}$$

is the statically screened Coulomb interaction between the carriers. The Coulomb sum in Eq. (10.1) includes all pair-wise interactions among the carriers, while it excludes the self-interaction with $i = j$. Note that the factor $\frac{1}{2}$ removes the double counting problem. In this article, we adopt the notation where ε_0 is the permittivity in free space, ε is the dielectric (screening) constant, and $|e|$ denotes the magnitude of the elementary charge where the electron has a charge $Q = -|e|$. Since quantum properties result from the operator character, we use form $\hat{O}$, i.e., we identify operators with a “hat” whenever it is not self-evident.

The Hamiltonian (10.1) shows that the carriers are coupled to the transversal field via the vector potential $\hat{\mathbf{A}}(\mathbf{r})$. As long as we do not have external longitudinal fields the system Hamiltonian does not contain any additional potential terms. In this case, it is convenient to adopt the Coulomb gauge

$$\nabla \cdot \hat{\mathbf{A}}(\mathbf{r}) = 0. \tag{10.3}$$

As a result, $\hat{\mathbf{p}} \cdot \hat{\mathbf{A}}(\mathbf{r}) = \hat{\mathbf{A}}(\mathbf{r}) \cdot \hat{\mathbf{p}}$ such that we can rewrite Eq. (10.1):

$$\hat{H}_N = \sum_{j=1}^{N} \left\{ \frac{\hat{\mathbf{p}}_j^2}{2m_0} + V_{\mathrm{L}}(\mathbf{r}_j) \right\} - \frac{Q}{m_0} \hat{\mathbf{A}}(\mathbf{r}_j) \cdot \hat{\mathbf{p}}_j + \frac{Q^2}{2m_0} \hat{\mathbf{A}}^2(\mathbf{r}_j) \\ + \frac{1}{2} \sum_{i \neq j}^{N} V(|\mathbf{r}_i - \mathbf{r}_j|) + \hat{H}_{\mathrm{em}} + \hat{H}_{\mathrm{ph}}. \tag{10.4}$$

Since we may want to treat a quantized light field, we include the free-field part $\hat{H}_{\mathrm{em}}$ and if we want to treat the interaction with lattice vibrations we also keep $\hat{H}_{\mathrm{ph}}$. This procedure introduces photon quanta for the light field and phonons for the vibrations. These quasi-particles and their contributions will be discussed later when the formalism of the second quantization is introduced.

We may now try to solve the many-body problem by starting from the Schrödinger equation

$$i\hbar \frac{\partial}{\partial t} \Psi(\mathbf{r}_1, .., \mathbf{r}_N) = \hat{H}_N \Psi(\mathbf{r}_1, .., \mathbf{r}_N), \tag{10.5}$$

if the light field is treated classically. For quantized light fields and lattice vibrations, the wave function contains additional coordinates related to photon and phonon degrees of freedom. Since electrons are indistinguishable Fermions, symmetry requirements demand that the exchange of any two-electron coordinates produces the same wave function with an opposite sign. Similarly, photons (phonons) among themselves are indistinguishable but they obey Bosonic statistics without the sign change.

We observe from the structure of Eq. (10.4) that $\hat{H}_N$ contains single-particle terms like $\hat{\mathbf{p}}^2/2m_0$ and two-particle interaction terms like $V(|\mathbf{r}_i - \mathbf{r}_j|)$. Even though the general form in Eq. (10.4) looks deceivingly simple, the two-particle terms lead to a genuine many-body problem where the solutions of the Schrödinger equation depend on each particle, photon, and phonon coordinate in a non-trivial manner such that beyond formal expressions, analytic solutions are generally not possible.

If a numerical solution of Eq. (10.5) is pursued, one typically discretizes each coordinate space $\mathbf{r}_j$ into M small intervals or volume units. With this straightforward procedure, $\Psi(\mathbf{r}_1, .., \mathbf{r}_N)$ can be presented numerically by an M^N-dimensional super matrix with complex-valued elements. If we use a modest estimate of $M = 1000$ discretizations, we find that the dimension of the matrix exceeds a computationally reasonable size already if N becomes larger than four. Since a typical semiconductor can easily have 10^{18} or more optically excited electrons within a cm^{-3}, direct solutions of Eq. (10.5) are impossible to obtain for realistic situations. Needless to say that the exact eigenstates or dynamics of the many-body Schrödinger equation are still very much unknown. However, one can develop and utilize sophisticated methods to generate

consistent approximative approaches to treat the many-body problem such that one can systematically improve the solutions.

10.2.2 Electrons in the Periodic Lattice Potential

As a starting point of our further many-body investigations, we first determine the basic characteristics of the non-interacting electrons in a periodic lattice potential. Since these effects are discussed in many solid-state theory textbooks, we focus here only on the central aspects that are relevant for our quantum-optical semiconductor theory. To compute the electronic eigenstates of the non-interacting electrons, we only need the single-particle part of Eq. (10.4). This leads to the eigenvalue problem

$$\left[\frac{\hat{\mathbf{p}}^2}{2m_0} + V_\mathrm{L}(\mathbf{r})\right]\varphi_{\lambda,\mathbf{k}}(\mathbf{r}) = \epsilon^\lambda_\mathbf{k}\varphi_{\lambda,\mathbf{k}}(\mathbf{r}), \tag{10.6}$$

where λ denotes a discrete set of states while $\mathbf{k}$ denotes the continuum of (quasi-) momentum states. The eigenfunctions $\varphi_{\lambda,\mathbf{k}(\mathbf{r})}$ are orthogonal, i.e.,

$$\int_{\mathsf{L}^3} \mathrm{d}^3r\varphi^*_{\lambda,\mathbf{k}}(\mathbf{r})\varphi_{\lambda',\mathbf{k}'}(\mathbf{r}) = \delta_{\lambda,\lambda'}\delta_{\mathbf{k},\mathbf{k}'}, \tag{10.7}$$

where L^3 is the quantization volume.

Physically, the lattice potential $V_\mathrm{L}(\mathbf{r})$ is the superposition of the attractive Coulomb interactions between an active electron and all the ions at their different lattice sites. Asit turns out, for our considerations we never need the explicit form of $V_\mathrm{L}(\mathbf{r})$. We only make use of some general features, such as the symmetry and periodicity properties of the potential, which reflect the structure of the crystal lattice. The periodicity of the effective lattice potential is expressed by the translational symmetry

$$V_\mathrm{L}(\mathbf{r}) = V_\mathrm{L}(\mathbf{r} + \mathbf{R}_n), \tag{10.8}$$

where $\mathbf{R}_n$ is a *lattice vector*, i.e., a vector that connects two identical sites in an infinite lattice. Since $V_\mathrm{L}(\mathbf{r})$ is periodic, the entire volume L^3 can be subdivided into identical *unit cells*. If this is done, the positions $\mathbf{r}$ and $\mathbf{r} + \mathbf{R}_n$ are n unit cells apart. Thus, it is convenient to expand the lattice vectors according to

$$\mathbf{R}_n = \sum_i n_i\mathbf{a}_i, \tag{10.9}$$

where n_i are integers and $\mathbf{a_i}$ are the basis vectors which span the unit cells. Note that the basis vectors are usually not unit vectors and they are generally not even orthogonal. The basis vectors point to the directions of the three axes of the unit

cell, which may have, e.g., a rhombic or more complicated shape. The basis vectors are parallel to the usual Cartesian unit vectors only in the case of orthogonal lattices such as the cubic one.

The specific symmetry of $V_{\mathrm{L}}(\mathbf{r})$ implies restrictions also for $\varphi_{\lambda,\mathbf{k}}(\mathbf{r})$. This symmetry requirement is known as the *Bloch theorem*

$$\varphi_{\lambda,\mathbf{k}}(\mathbf{r}+\mathbf{R}_n) = \mathrm{e}^{\,\mathrm{i}\,\mathbf{k}\cdot\mathbf{R}_n}\varphi_{\lambda,\mathbf{k}}(\mathbf{r}), \tag{10.10}$$

which states that a translation by $\mathbf{R}_n$ can only result in a phase shift $\mathrm{e}^{\,i\,\mathbf{k}\cdot\mathbf{R}_n}$ of the original wave function.

To satisfy the Bloch theorem, we make the ansatz

$$\varphi_{\lambda,\mathbf{k}}(\mathbf{r}) = \frac{\mathrm{e}^{\,\mathrm{i}\,\mathbf{k}\cdot\mathbf{r}}}{\mathsf{L}^{3/2}}u_{\lambda,\mathbf{k}}(\mathbf{r})\,. \tag{10.11}$$

Here, $u_{\lambda,\mathbf{k}}(\mathbf{r})$ is the *Bloch function*. The ansatz (10.11) fulfills the Bloch theorem (10.10) if u_{λ} is periodic in real space:

$$u_{\lambda,\mathbf{k}}(\mathbf{r}) = u_{\lambda,\mathbf{k}}(\mathbf{r}+\mathbf{R}_n), \tag{10.12}$$

i.e., if the Bloch function has the lattice periodicity. With the help of the basic property of the momentum operator, $\hat{\mathbf{p}}\,\mathrm{e}^{\,\mathrm{i}\,\mathbf{k}\cdot\mathbf{r}} = \hbar\mathbf{k}\,\mathrm{e}^{\,\mathrm{i}\,\mathbf{k}\cdot\mathbf{r}}$, we find

$$\begin{aligned} L^{3/2}\hat{\mathbf{p}}^2\varphi_{\lambda,\mathbf{k}}(\mathbf{r}) &= \hat{\mathbf{p}}\cdot\hat{\mathbf{p}}\left[\mathrm{e}^{\,\mathrm{i}\,\mathbf{k}\cdot\mathbf{r}}u_{\lambda,\mathbf{k}}(\mathbf{r})\right] \\ &= \hat{\mathbf{p}}\cdot\left[\hbar\mathbf{k}\,\mathrm{e}^{\,\mathrm{i}\,\mathbf{k}\cdot\mathbf{r}}u_{\lambda,\mathbf{k}}(\mathbf{r}) + \mathrm{e}^{\,\mathrm{i}\,\mathbf{k}\cdot\mathbf{r}}\,\hat{\mathbf{p}}u_{\lambda,\mathbf{k}}(\mathbf{r})\right] \\ &= \mathrm{e}^{\,\mathrm{i}\,\mathbf{k}\cdot\mathbf{r}}\left[\hbar^2\mathbf{k}^2u_{\lambda,\mathbf{k}}(\mathbf{r}) + 2\hbar\mathbf{k}\cdot\hat{\mathbf{p}}\,u_{\lambda,\mathbf{k}}(\mathbf{r})\right. \\ &\quad\left. + \hat{\mathbf{p}}^2u_{\lambda,\mathbf{k}}(\mathbf{r})\right]. \end{aligned} \tag{10.13}$$

Inserting this result into Eq. (10.6), we obtain

$$\left[\frac{\hat{\mathbf{p}}^2}{2m_0} + \frac{\hbar}{m_0}\mathbf{k}\cdot\hat{\mathbf{p}} + V_{\mathrm{L}}(\mathbf{r})\right]u_{\lambda,\mathbf{k}}(\mathbf{r}) = \left[\epsilon^{\lambda}_{\mathbf{k}} - \frac{\hbar^2\mathbf{k}^2}{2m_0}\right]u_{\lambda,\mathbf{k}}(\mathbf{r}), \tag{10.14}$$

which will be the starting point for the $\mathbf{k}\cdot\mathbf{p}$ analysis. Once $u_{\lambda,\mathbf{k}}(\mathbf{r})$ is determined, the solution of the original Eq. (10.6) is directly obtained by using Eq. (10.11).

10.2.3 k·p Theory

In this section, we describe those aspects of the $\mathbf{k}\cdot\mathbf{p}$ perturbation theory that we need in later derivations and which allow us to discuss qualitative properties of the band structure. The basic idea behind this approximation is to assume that

one has solved the band structure problem at some point $\mathbf{k}_0$ with high symmetry. Here, we will take this point as $\mathbf{k}_0 = 0$, which is called the Γ-point. In particular, we assume that we know all energy eigenvalues ϵ_0^λ and the corresponding Bloch functions $u_{\lambda,\mathbf{k}_0=0}(\mathbf{r}) = u_{\lambda,0}(\mathbf{r})$. For the following manipulations, we now adopt Dirac's abstract notation with $|\lambda,\mathbf{k}\rangle$ and $|\lambda\rangle \equiv |\mathbf{0},\lambda\rangle$, which both have the usual real-space representation $u_{\lambda,\mathbf{k}}(\mathbf{r}) = \langle\mathbf{r}|\lambda,\mathbf{k}\rangle$.

In order to compute the Bloch functions $|\lambda,\mathbf{k}\rangle$ and the corresponding energy eigenvalues $\epsilon_\mathbf{k}^\lambda$ for $\mathbf{k}$ in the vicinity of the Γ-point, we expand the lattice periodic function $|\lambda,\mathbf{k}\rangle$ in terms of the known functions $|\lambda\rangle$ which form a complete set. We rewrite Eq. (10.14) as

$$\left[\hat{H}_0 + \frac{\hbar}{m_0}\mathbf{k}\cdot\hat{\mathbf{p}}\right]|\lambda,\mathbf{k}\rangle = \left[\epsilon_\mathbf{k}^\lambda - \frac{\hbar^2\mathbf{k}^2}{2m_0}\right]|\lambda,\mathbf{k}\rangle, \tag{10.15}$$

where

$$\hat{H}_0 = \frac{\hat{\mathbf{p}}^2}{2m_0} + \hat{V}_\mathrm{L}. \tag{10.16}$$

The idea now is to treat the $\mathbf{k}\cdot\mathbf{p}$ term as a perturbation to the Hamiltonian $\hat{H}_0$. Since we assume that the eigenvalue problem $\hat{H}_0|\lambda\rangle = \epsilon_\mathbf{0}^\lambda|\lambda\rangle$ is known, we derive a perturbative solution for $|\lambda,\mathbf{k}\rangle$. In general, degenerate perturbation theory is needed if several bands are degenerate at the Γ-point. Here, we restrict ourselves to the simpler case of non-degenerate perturbation theory for notational simplicity. For the conduction band, this approach is exact, and it can be generalized if more than one valence band shall be considered [4].

Using general parity arguments, we see that

$$\langle\lambda|\hat{\mathbf{p}}|\lambda\rangle = 0, \tag{10.17}$$

i.e., there is no first-order energy correction to $\epsilon_\mathbf{k}^\lambda$. Thus, we have to apply at least second-order non-degenerate perturbation theory to obtain

$$|\mathbf{k},\lambda\rangle = |\lambda\rangle + \frac{\hbar}{m_0}\sum_{\eta\neq\lambda}\frac{|\eta\rangle\mathbf{k}\cdot\langle\eta|\mathbf{p}|\lambda\rangle}{\epsilon_0^\lambda - \epsilon_0^\eta} + \mathrm{O}(\mathbf{k}^2) \tag{10.18}$$

and

$$\epsilon_\mathbf{k}^\lambda = \epsilon_0^\lambda + \frac{\hbar^2\mathbf{k}^2}{2m_0} + \sum_{\eta\neq\lambda}\frac{\hbar^2}{m_0^2}\frac{(\mathbf{k}\cdot\langle\lambda|\mathbf{p}|\eta\rangle)(\mathbf{k}\cdot\langle\eta|\mathbf{p}|\lambda\rangle)}{\epsilon_0^\lambda - \epsilon_0^\eta} + \mathrm{O}(\mathbf{k}^3). \tag{10.19}$$

These results become increasingly accurate for sufficiently small $\mathbf{k}$.

In order to gain insight into the $\mathbf{k} \cdot \mathbf{p}$ results, we consider the simplest case with two discrete states $|0\rangle$ and $|1\rangle$. Using Cartesian coordinates with $\mathbf{p} = (p_1, p_2, p_3)$, we find

$$\epsilon_{\mathbf{k}}^1 = \epsilon_0^1 + \frac{\hbar^2 \mathbf{k}^2}{2m_0} + \sum_{i,j=1}^{3} \frac{\hbar^2 k_i k_j}{2m_0} \frac{2\mathsf{p}_i^* \mathsf{p}_j}{m_0 E_g} \tag{10.20}$$

and

$$\epsilon_{\mathbf{k}}^0 = \epsilon_0^0 + \frac{\hbar^2 \mathbf{k}^2}{2m_0} - \sum_{i,j=1}^{3} \frac{\hbar^2 k_i k_j}{2m_0} \frac{2\mathsf{p}_i^* \mathsf{p}_j}{m_0 E_g}, \tag{10.21}$$

where we define the unrenormalized band gap $E_g = \epsilon_0^1 - \epsilon_0^0$ and the momentum matrix element $\mathsf{p}_j = \langle 0|p_j|1\rangle$. Since the energy has a quadratic $\mathbf{k}$-dependence, it is meaningful to introduce the effective mass tensor:

$$\left(\frac{1}{m_{\text{eff}}}\right)_{ij} = \frac{1}{m_0}\left(\delta_{ij} \pm \frac{2\mathsf{p}_i^* \mathsf{p}_j}{m_0 E_g}\right). \tag{10.22}$$

In isotropic cases, such as in cubic lattice symmetry, the effective masses are scalar quantities:

$$m_c = \frac{m_0}{1 + \frac{2\mathsf{p}^2}{m_0 E_g}} \tag{10.23}$$

for the upper level $|1\rangle$ and

$$m_v = \frac{m_0}{1 - \frac{2\mathsf{p}^2}{m_0 E_g}} \tag{10.24}$$

for the lower level $|0\rangle$. In this situation, the $\mathbf{k} \cdot \mathbf{p}$ energies become

$$\epsilon_{\mathbf{k}}^c = \epsilon_0^1 + \frac{\hbar^2 \mathbf{k}^2}{2m_e} \tag{10.25}$$

and

$$\epsilon_{\mathbf{k}}^v = \epsilon_0^0 + \frac{\hbar^2 \mathbf{k}^2}{2m_v}, \tag{10.26}$$

where the upper level is called *conduction band* ($1 \equiv c$) and the lower level is known as *valence band* ($0 \equiv v$). By starting from Eq. (10.19), one obtains a more general isotropic effective mass for the band λ:

$$\frac{1}{m_\lambda} = \frac{1}{m_0} + \frac{2}{m_0^2} \sum_{\eta \neq \lambda} \frac{\langle \lambda|p|\eta\rangle \langle \eta|p|\lambda\rangle}{\epsilon_0^\lambda - \epsilon_0^\eta} \tag{10.27}$$

and

$$\epsilon_{\mathbf{k}}^{\lambda} = \epsilon_0^{\lambda} + \frac{\hbar^2 \mathbf{k}^2}{2m_{\lambda}}, \tag{10.28}$$

where we have used an isotropic approximation

$$\mathbf{k} \cdot \langle \lambda | \mathbf{p} | \eta \rangle \mathbf{k} \cdot \langle \eta | \mathbf{p} | \lambda \rangle = \mathbf{k}^2 \langle \lambda | p | \eta \rangle \langle \eta | p | \lambda \rangle. \tag{10.29}$$

The non-isotropic generalization of Eq. (10.29) leads to an effective mass tensor in analogy to Eq. (10.22); however, we concentrate here on isotropic systems.

For a sufficiently large momentum matrix element, the effective mass of the valence band usually is negative while the effective mass of the conduction band is positive and much smaller than the free-electron mass. Equations (10.23) and (10.24) show that the effective masses are determined by the interband matrix element of the momentum operator and by the energy gap. Once $m_{\mathrm{e}} = m_c$ and $m_{\mathrm{h}} = -m_v$ are known, we may define the reduced electron–hole mass m_{r}:

$$\frac{1}{m_{\mathrm{r}}} = \frac{1}{m_{\mathrm{e}}} + \frac{1}{m_{\mathrm{h}}} = \frac{4\mathsf{p}^2}{m_0^2 E_{\mathrm{g}}}, \tag{10.30}$$

which follows directly from Eqs. (10.23) and (10.24). This result is often used to estimate the value of p^2.

To illustrate a typical band structure used in many semiconductor quantum-optical investigations, we consider $\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{As}/\mathrm{GaAs}$ quantum-well systems, where thin layers of $\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{As}$ wells are sandwiched between GaAs barriers. $\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{As}$ is a compound semiconductor where the subscript x denotes the relative percentage of gallium atoms that have been replaced by indium atoms. For such structures, the classical and quantum optics take place close to the direct band gap which is roughly 1.5 eV wide. With suitable indium concentrations, the system becomes effectively a non-degenerate two-band system which has a conduction band with $m_c = +0.0665 m_0$ and valence band with $m_v = -0.235 m_0$. The corresponding band structure is sketched in Fig. 10.1 comparing a parabolic approximation (solid line) with the result of a full band structure computation (dashed line).

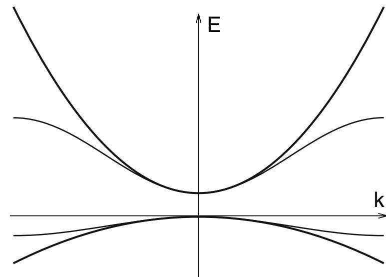


Fig. 10.1 Schematic sketch of semiconductor band structure with two bands, compared to the parabolic bands using an effective mass appproximation. The mass is determined by the band curvature at the Γ-point $k = 0$. From Haug and Koch [4]

10.2.4 Second Quantization of the Carrier System

Even though the exact many-body wave function is unknown, we can always construct a complete basis set from products of single-particle wave functions in the properly anti-symmetrized form. For example, the Slater determinants represent conveniently anti-symmetrized states constructed from N known single-particle wave functions. If the particle number increases, the number of combinations becomes enormous, which implies a difficult book-keeping problem. To avoid these difficulties, one often introduces an equivalent formalism, where one can create or annihilate a particle in any desired single-particle state such that the resulting wave function has the correct Fermi anti-symmetry. This procedure can be obtained either from the occupation-number representation for identical particles or from the so-called second quantization [5, 6]. While the former introduces the creation and annihilation operators as convenient operators in order to create many-body states obeying the correct symmetry requirements, the name "second quantization" stems from an alternative derivation in which the single-particle wave function is considered as a "classical" field and field quantization is applied very much the same as for the quantization of the electromagnetic field [3]. In the following discussion, we present applications to our semiconductor quantum-optical problem.

The many-body properties of the carriers are determined by the second quantization field operators:

$$\hat{\Psi}(\mathbf{r}) = \sum_{\lambda,\mathbf{k}} a_{\lambda,\mathbf{k}}\ \varphi_{\lambda,\mathbf{k}}(\mathbf{r}), \tag{10.31}$$

where the operator $a_{\lambda,\mathbf{k}}$ annihilates an electron with momentum $\mathbf{k}$ in the state λ, which combines the band and the spin index. The corresponding single-particle wave functions, $\varphi_{\lambda,\mathbf{k}}(\mathbf{r})$, are orthogonal and form a complete set. Since carriers are Fermions, the operators $a_{\lambda,\mathbf{k}}$ obey anti-commutation relations:

$$\left[a_{\lambda,\mathbf{k}}, a^{\dagger}_{\lambda',\mathbf{k}'}\right]_{+} = a_{\lambda,\mathbf{k}}, a^{\dagger}_{\lambda',\mathbf{k}'} + a^{\dagger}_{\lambda',\mathbf{k}'}, a_{\lambda,\mathbf{k}} = \delta_{\mathbf{k},\mathbf{k}'}\delta_{\lambda,\lambda'}, \tag{10.32}$$

$$\left[a^{\dagger}_{\lambda,\mathbf{k}}, a^{\dagger}_{\lambda',\mathbf{k}'}\right]_{+} = \left[a_{\lambda,\mathbf{k}}, a_{\lambda',\mathbf{k}'}\right]_{+} = 0. \tag{10.33}$$

The second quantized form of a single-particle operator $O_1(\mathbf{r})$ is

$$\begin{aligned}\hat{O}_1 &= \int \hat{\Psi}^{\dagger}(\mathbf{r}) O_1(\mathbf{r}) \hat{\Psi}(\mathbf{r}) \mathrm{d}^3 r \\ &= \sum_{\mathbf{k},\mathbf{k}',\lambda,\lambda'} a^{\dagger}_{\lambda,\mathbf{k}} a_{\lambda',\mathbf{k}'} \int \varphi^{*}_{\lambda,\mathbf{k}}(\mathbf{r}) O_1(\mathbf{r}) \varphi_{\lambda',\mathbf{k}'} \mathrm{d}^3 r,\end{aligned} \tag{10.34}$$

and the second quantized form of a two-particle operator $O_2(\mathbf{r},\mathbf{r}')$ is

$$\begin{aligned}\hat{O}_2 &= \int \hat{\Psi}^\dagger(\mathbf{r})\hat{\Psi}^\dagger(\mathbf{r}')O_1(\mathbf{r},\mathbf{r}')\hat{\Psi}(\mathbf{r}')\hat{\Psi}(\mathbf{r})\mathrm{d}^3r'\mathrm{d}^3r \\ &= \sum_{\mathbf{k},\mathbf{k}',\mathbf{p},\mathbf{p}'}\ \sum_{\lambda,\lambda',\nu,\nu'} a^\dagger_{\lambda,\mathbf{k}}a^\dagger_{\nu,\mathbf{p}}a_{\nu',\mathbf{p}'}a_{\lambda',\mathbf{k}'} \\ &\quad\times \int \varphi^*_{\lambda,\mathbf{k}}(\mathbf{r})\varphi^*_{\nu,\mathbf{p}}(\mathbf{r}')\ O_2(\mathbf{r},\mathbf{r}')\ \varphi_{\nu',\mathbf{p}'}(\mathbf{r}')\varphi_{\lambda',\mathbf{k}'}(\mathbf{r})\mathrm{d}^3r'\mathrm{d}^3r.\end{aligned} \tag{10.35}$$

We see that besides the electron creation and annihilation operator only matrix elements between the single-particle wave functions enter into the theory.

In principle, one can choose any complete set of single-particle wave functions $\varphi_{\lambda,\mathbf{k}}$. In most of the cases, it is convenient to choose the orthogonal basis of Bloch functions which diagonalizes the non-interacting electron Hamiltonian:

$$\left[\frac{\hat{\mathbf{p}}^2}{2m_0} + V_\mathrm{L}(\mathbf{r})\right]\varphi_{\lambda,\mathbf{k}}(\mathbf{r}) = \epsilon^\lambda_\mathbf{k}\varphi_{\lambda,\mathbf{k}}(\mathbf{r}). \tag{10.36}$$

Thus, we describe the many-body system in terms of Bloch electrons. In particular, we will use the $\mathbf{k}\cdot\mathbf{p}$ wave functions $\varphi_{\lambda,\mathbf{k}}(\mathbf{r})$, which we assume to be explicitly known. Starting from Eq. (10.4), we find that the non-interacting electron Hamiltonian follows from

$$\begin{aligned}\hat{H}_0 &= \int \hat{\Psi}^\dagger(\mathbf{r})\left[\frac{\hat{\mathbf{p}}^2}{2m_0} + V_\mathrm{L}(\mathbf{r})\right]\hat{\Psi}(\mathbf{r})\mathrm{d}^3r \\ &= \sum_{\lambda,\mathbf{k},\lambda',\mathbf{k}'} a^\dagger_{\lambda,\mathbf{k}}a_{\lambda',\mathbf{k}'}\int \varphi^*_{\lambda,\mathbf{k}}(\mathbf{r})\left[\frac{\mathbf{p}^2}{2m_0} + V_\mathrm{L}(\mathbf{r})\right]\varphi_{\lambda',\mathbf{k}'}(\mathbf{r})\mathrm{d}^3r \\ &= \sum_{\lambda,\mathbf{k},\lambda',\mathbf{k}'} a^\dagger_{\lambda,\mathbf{k}}a_{\lambda',\mathbf{k}'}\int \varphi^*_{\lambda,\mathbf{k}}(\mathbf{r})\epsilon^{\lambda'}_{\mathbf{k}'}\varphi_{\lambda',\mathbf{k}'}(\mathbf{r})\mathrm{d}^3r \\ &= \sum_{\lambda,\mathbf{k}} \epsilon^\lambda_\mathbf{k} a^\dagger_{\lambda,\mathbf{k}}a_{\lambda,\mathbf{k}},\end{aligned} \tag{10.37}$$

where we have used Eq. (10.36) and the orthogonality of $\varphi_{\lambda,\mathbf{k}}(\mathbf{r})$. Equation (10.37) shows that the non-interacting part yields the Hamiltonian of a simple harmonic oscillator if the electrons are presented in the Bloch basis. Besides $\hat{H}_0$, we clearly need to determine also the Coulomb interaction among the Bloch electrons and the coupling between electron and quantized light field and lattice vibrations. Before entering into this analysis in Section 10.3.5, we generalize the description of active Bloch electrons beyond the three-dimensional bulk case in the next section and introduce the quantization of electromagnetic fields and lattice vibrations in the remainder of Section 10.2.

10.2.5 Systems with Reduced Effective Dimensionality

Several crystal growth techniques allow the grower to prepare semiconductor samples where one periodic lattice is changed to another one by alternating chemical compounds in different growth layers. These manufacturing technologies have reached a quality level where the layer interfaces can be controlled with atomic accuracy. If planar structures are grown, e.g., in z-direction, the lattice periodic potential in Eq. (10.36) then has to be replaced by

$$V_{\mathrm{L}}(\mathbf{r}) = V_{\mathrm{L}}^{i}(\mathbf{r}), \quad \text{for each} \quad z_i < z < z_{i+1}, \tag{10.38}$$

where z_i indicates the positions of the different interfaces. Within each interval $z_i < z < z_{i+1}$, the lattice periodic potential depends on the chemical compounds in that region. This additional feature complicates the procedure to find the exact solution $\varphi_{\lambda,\mathbf{k}}(\mathbf{r})$ of Eq. (10.36). However, most of the relevant results can be obtained by using an approximative approach that makes use of the fact that the multilayer structures are mesoscopic, i.e., large in comparison to the microscopic atomic scale but small in comparison to the overall sample dimensions. In other words, the active layers have a thickness $L_c^i = z_{i+1} - z_i$, which is much wider than the lattice unit cell while L_c^i is much smaller than the macroscopic sample size. If the mesoscopic layer thickness L_c^i exceeds a few unit cells, one finds a well-defined band structure within each layer. The band structures in the different layers can be assumed to be the bulk band structures shifted by the respective confinement energy levels for electrons and holes.

To analyze the fundamental confinement effects, we consider a structure where one mesoscopic planar layer L_c is sandwiched between two identical bulk barriers. Furthermore, we assume that the mesoscopic layer has much lower ϵ_0^c than the surrounding bulk. This construction is known as quantum well since electrons tend to be trapped in the region with the smallest ϵ_0^c. For planar quantum-well systems, it is useful to separate the three-dimensional space coordinate $\mathbf{r} = (\mathbf{r}_\parallel, z)$ into a two-dimensional vector $\mathbf{r}_\parallel$ in the quantum-well plane and the one-dimensional coordinate z perpendicular to the quantum well. This system is clearly fully periodic within the quantum-well plane such that we may apply the Bloch theorem and ansatz (10.11) for the planar dependency. However, the original ansatz has to be modified to include the actual z-dependence. If the chemical compounds are not considerably different, the lattice periodic Bloch function u can be assumed to be the same throughout the sample since it depends on the microscopic scale. However, the different layers can be assumed to have clearly different ϵ_0^c, which is well defined due to the mesoscopic size of L_c. In this situation, the quantum-well confinement modifies only the z-dependent part of the envelope function. Thus, we may introduce the *envelope-function approximation*

$$\varphi_{\lambda,\mathbf{k}_\parallel,n}(\mathbf{r}) = \xi_{\lambda,n}(z)\frac{1}{\sqrt{S}}\mathrm{e}^{\,\mathrm{i}\mathbf{k}_\parallel\cdot\mathbf{r}_\parallel}u_{\lambda,\mathbf{k}_\parallel}(\mathbf{r}), \tag{10.39}$$

where $\xi_{\lambda,n}(z)$ is the mesoscopic confinement wave function for the level n, $\mathbf{k}_{\|}$ is the carrier momentum in the quantum-well plane, and $u_{\lambda,\mathbf{k}_{\|}}(\mathbf{r},z)$ is the lattice periodic Bloch function. Since $\xi_{\lambda,n}(z)$ describes a mesoscopic envelope, it is affected by the z-dependency of the band structure. Within the effective mass approximation, we find

$$\left[-\frac{\hbar^2}{2m_\lambda}\frac{\partial^2}{\partial z^2}+V^\lambda_{\text{conf}}(z)\right]\xi_{\lambda,n}(z)=\epsilon_0^{\lambda,n}\xi_{\lambda,n}(z), \tag{10.40}$$

where $V^\lambda_{\text{conf}}(z)$ is the confinement potential determined by the z-dependent changes in the effective ϵ^λ_k within each quantum-well layer. The eigenenergy $\epsilon_0^{\lambda,n}$ defines the zero level of the $\mathbf{k}_{\|}$-dependent energy of the Bloch electrons:

$$\epsilon^{\lambda,n}_{\mathbf{k}_{\|}}=\epsilon_0^{\lambda,n}+\frac{\hbar^2k_{\|}^2}{2m_\lambda}, \tag{10.41}$$

where we have used the effective mass approximation. Even though the envelope-function approximation is not an exact solution of Eq. (10.36) with the potential (10.38), it usually is reasonably accurate and can be used to describe most quantum-well structures.

To demonstrate the principal effects of the quantum-well confinement, we consider a case where the quantum confinement is very strong, actually infinite:

$$V^\lambda_{\text{conf}}(z)=\begin{cases}\epsilon^\lambda_0, & |z|<L_c/2\\ \infty, & |z|>L_c/2\,.\end{cases} \tag{10.42}$$

In this situation, Eq. (10.40) represents the usual particle-in-a-box problem which has the eigenfunctions

$$\xi_n(z)=\begin{cases}\sqrt{\frac{2}{L_c}}\sin\frac{\pi}{L_c}n(z+L_c/2) & |z|<L_c/2\\ 0, & |z|>L_c/2,\end{cases} \tag{10.43}$$

and the eigenenergies (subbands)

$$\epsilon_0^{\lambda,n}=\epsilon_0^\lambda+\frac{\pi^2\hbar^2}{2m_\lambda L_c^2}n^2, \tag{10.44}$$

where $n=1,2,\ldots$. As the size of L_c is reduced, the energy differences between the different confinement levels n increase proportionally to L_c^{-2}. Thus, each subband becomes well separated for small enough L_c such that it is easy to configure a situation where the light field excitation and the subsequent many-body dynamics involves only the lowest confinement level even at room temperature.

As a consequence of the confinement, the lowest energies of different bands can be shifted, see Eq. (10.44). This property can be used to a certain extent to tune the resonance energies to a desired range. In addition, if the original three-dimensional band structure has degenerate bands in the vicinity of the optical transitions, it is rather simple in quantum-well structures to remove this degeneracy either via the confinement effects or by introducing some strain. Thus, one can design semiconductors that effectively behave like two-band systems. In this article, we mainly consider such systems as representative examples. More difficult band structures can be treated as well however, this requires additional book keeping of the band indices and increased numerical complexity.

We write the Bloch wavefunction as

$$\varphi_{c,\mathbf{k}_\parallel}(\mathbf{r}) = \xi(z)\frac{1}{\sqrt{L^2}}\mathrm{e}^{\mathrm{i}\mathbf{k}_\parallel\cdot\mathbf{r}_\parallel}u_{c,\mathbf{k}_\parallel}(\mathbf{r}) \tag{10.45}$$

for the conduction-band electrons and

$$\varphi_{v,\mathbf{k}_\parallel}(\mathbf{r}) = \xi(z)\frac{1}{\sqrt{L^2}}\mathrm{e}^{\mathrm{i}\mathbf{k}_\parallel\cdot\mathbf{r}_\parallel}u_{v,\mathbf{k}_\parallel}(\mathbf{r}) \tag{10.46}$$

for the valence-band electrons. Since the electrons are in the lowest confinement level, we omit the subband index n. The corresponding wave functions according to Eq. (10.43) are

$$\xi(z) = \sqrt{\frac{2}{L_c}}\cos\frac{\pi}{L_c}z, \quad |z| < L_c/2, \tag{10.47}$$

and the energies can be written as

$$\epsilon^c_{\mathbf{k}_\parallel} = \epsilon^c_0 + \frac{\hbar^2\mathbf{k}^2_\parallel}{2m_\mathrm{e}}, \tag{10.48}$$

$$\epsilon^c_{\mathbf{k}_\parallel} = \epsilon^v_0 - \frac{\hbar^2\mathbf{k}^2_\parallel}{2m_\mathrm{h}}, \tag{10.49}$$

where the effective mass approximation has been applied. By repeating the derivation that for bulk systems leads to Eq. (10.37), we find the Hamiltonian for the non-interacting quantum-well electrons

$$\hat{H}_0 = \sum_{\mathbf{k}_\parallel}\left(\epsilon^c_{\mathbf{k}_\parallel}a^\dagger_{c,\mathbf{k}_\parallel}a_{c,\mathbf{k}_\parallel} + \epsilon^v_{\mathbf{k}_\parallel}a^\dagger_{v,\mathbf{k}_\parallel}a_{v,\mathbf{k}_\parallel}\right). \tag{10.50}$$

The semiconductor system can also be confined in more than one direction leading to a further reduction of the effective system dimensionality. The

confinement in two directions yields the one-dimensional so-called quantum-wire structures, while the completely confined structures are known as effectively zero-dimensional quantum dots.

For quantum wires, it is once again useful to separate $\mathbf{r} = (\mathbf{r}_{\|}, z)$ where now $\mathbf{r}_{\|}$ denotes the confinement directions. For this situation, the envelope-function approximation is

$$\varphi_{c,k_z}(\mathbf{r}) = \xi(\mathbf{r}_{\|}) \frac{1}{\sqrt{L}} \mathrm{e}^{\mathrm{i}k_z z} u_{c,k_z}(\mathbf{r}) \tag{10.51}$$

for the conduction-band electrons and

$$\varphi_{v,k_z}(\mathbf{r}) = \xi(\mathbf{r}_{\|}) \frac{1}{\sqrt{L}} \mathrm{e}^{\mathrm{i}k_z z} u_{v,k_z}(\mathbf{r}) \tag{10.52}$$

for the valence band, since we have assumed confinement to the lowest level of the two-band system in analogy to the quantum-well case. For the mesoscopic quantum wire, the confinement function can be calculated from

$$\left[-\frac{\hbar^2}{2m_\lambda} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) + V^{\lambda}_{\mathrm{conf}}(x, y) \right] \xi_{\lambda,n}(x, y) = \epsilon_0^{\lambda,n} \xi_{\lambda,n}(x, y), \tag{10.53}$$

with $\mathbf{r}_{\|} = (x, y)$. By choosing a harmonic confinement, we find the lowest confinement level

$$\xi(\mathbf{r}_{\|}) = \sqrt{\pi} R^2 \mathrm{e}^{-\mathbf{r}_{\|}^2/R^2}, \tag{10.54}$$

where R defines the confinement scale.

In our numerical evaluations, we will often use a two-band quantum wire as a representative system to study quantum optical effects. The corresponding non-interacting part of the Hamiltonian is

$$\hat{H}_0 = \sum_{k_z} \left(\epsilon^c_{k_z} a^\dagger_{c,k_z} a_{c,k_z} + \epsilon^v_{k_z} a^\dagger_{v,k_z} a_{v,k_z} \right), \tag{10.55}$$

with the effective mass energies

$$\epsilon^c_{k_z} = \epsilon^c_0 + \frac{\hbar^2 k_z^2}{2m_\mathrm{e}}, \tag{10.56}$$

$$\epsilon^v_{k_z} = \epsilon^v_0 - \frac{\hbar^2 k_z^2}{2m_\mathrm{h}}, \tag{10.57}$$

in analogy to the quantum-well system.

If the semiconductor system is confined in all directions, we obtain quantum dots. In the envelope-function approximation, the corresponding Bloch wave-functions are

$$\varphi_{c,n}(\mathbf{r}) = \xi_{c,n}(\mathbf{r}) u_c(\mathbf{r}) \tag{10.58}$$

and

$$\varphi_{v,n}(\mathbf{r}) = \xi_{v,n}(\mathbf{r}) u_v(\mathbf{r}), \tag{10.59}$$

where n refers to the quantum number of

$$\left[-\frac{\hbar^2}{2m_\lambda} \nabla^2 + V^\lambda_{\text{conf}}(\mathbf{r}) \right] \xi_{\lambda,n}(\mathbf{r}) = \epsilon^{\lambda,n} \xi_{\lambda,n}(\mathbf{r}) \, . \tag{10.60}$$

In general, these solutions consist of discrete states bound inside the quantum dot, plus energetically higher unconfined states. Since the electrons can occupy each dot level only twice (once each for spin up and down), it is natural to include many confinement levels for dots even when the confinement is strong. The corresponding non-interacting Hamiltonian is then

$$\hat{H}_0 = \sum_n \left(\epsilon^c_n a^\dagger_{c,n} a_{c,n} + \epsilon^v_n a^\dagger_{v,n} a_{v,n} \right). \tag{10.61}$$

If the energy levels of the dot are well separated, the quantum-dot system allows for spectroscopy between discrete levels in analogy to atomic systems.

10.2.6 Electron Density of States

In the definition of the Bloch functions, we deliberately did not specify how the quantization volume L^d has to be chosen for effectively d-dimensional systems. Since all real samples have a different finite size, it is useful to assume that the sample consists of many identical parts with volume L^d. This way, we have the same quantization volume for all relevant systems if we implement periodic boundary conditions at each surface of L^d. As a result, the plane-wave parts of the envelope functions have to fulfill the condition

$$\mathrm{e}^{\mathrm{i}k_j \mathsf{L}} = 1 \quad \Leftrightarrow \quad k_j \mathsf{L} = 2\pi n, \tag{10.62}$$

where k_j is the Cartesian component of the wave vector. Thus, each k_j is discretized according to

$$k_j = \frac{2\pi}{\mathsf{L}} n \equiv n \Delta k, \tag{10.63}$$

which defines the momentum difference $\Delta k = \frac{2\pi}{\mathsf{L}}$.

For a large enough quantization volume, Δk becomes infinitesimal such that k_j becomes a continuous variable. However, by using a formally finite L together with the discretization (10.63), we can introduce an efficient way to handle sums over k_j, which occur quite frequently in our investigation. The typical form contains a generic function $F_{\mathbf{k}}$ in an expression

$$\begin{aligned} \frac{1}{\mathsf{L}^d}\sum_{\mathbf{k}} F_{\mathbf{k}} &= \frac{1}{(2\pi)^d}\left(\frac{2\pi}{\mathsf{L}}\right)^d \sum_{\mathbf{k}} F_{\mathbf{k}} \\ &= \frac{1}{(2\pi)^d}\sum_{\mathbf{k}} F_{\mathbf{k}}(\Delta k)^d \\ &= \frac{1}{(2\pi)^d}\int_{\mathsf{L}^d} F_{\mathbf{k}}\mathrm{d}^d k, \end{aligned} \tag{10.64}$$

where the last step follows for large L and infinitesimal Δk since then the second line becomes the standard definition of an integral. This property will be used several times in further derivations.

Many relevant integrals defined by Eq. (10.64) have an integrand which depends only on the magnitude $k = |\mathbf{k}|$. For these cases, it is convenient to perform the integration in either radial or spherical coordinates. The corresponding form of the integral (10.64) follows from

$$\frac{1}{\mathsf{L}^d}\sum_{\mathbf{k}} F_k = \frac{\Omega_d}{(2\pi)^d}\int_0^\infty k^{d-1} F_k \mathrm{d}k, \tag{10.65}$$

where $\Omega_{d=1} = 2$, $\Omega_{d=2} = 2\pi$, and $\Omega_{d=3} = 4\pi$ contain the integral over the angles in different dimensions. Since $F_{|k|} = F(E)$ is often known as function of energy $E = \frac{\hbar^2 k^2}{2m_\lambda}$, one may change the integration variables to obtain

$$\frac{1}{\mathsf{L}^d}\sum_{\mathbf{k}} F_{\mathbf{k}} = \int_0^\infty g_d(E)F(E)\mathrm{d}E. \tag{10.66}$$

Here, the quantity

$$g_d(E) = \frac{\Omega_d}{(2\pi)^d}\frac{1}{2}\left(\frac{2m_\lambda}{\hbar^2}\right)^{d/2} E^{d/2-1} \tag{10.67}$$

is known as the *energy density of states* for the particles λ. The functional form of $g_d(E)$ depends strongly on the effective system dimension. We will see later that this has profound consequences for physically measurable quantities.

10.2.7 Quantization of Electromagnetic Fields

In order to describe quantum-optical effects, we have to know how to treat the electromagnetic fields in a quantized form. Detailed derivations can be found in many quantum-optics textbooks, e.g., in Cohen-Tannoudji et al. [3], and we will only summarize the key steps here.

General starting point is the classical electro-magnetic field energy:

$$H_{\mathrm{em}} = \frac{\varepsilon_0}{2} \int \mathrm{d}^3 r \Big[n^2(\mathbf{r}) |\mathbf{E}_{\mathrm{T}}(\mathbf{r})|^2 + c^2 |\mathbf{B}(\mathbf{r})|^2 \Big], \tag{10.68}$$

where c is the vacuum speed of light given by $c = (\varepsilon_0 \mu_0)^{-1/2}$. The polarizability of the optically passive dielectric structure surrounding the active semiconductor material is described by a (possibly space-dependent) refractive index $n(\mathbf{r})$. In most relevant cases, such as dielectric Bragg mirrors, photonic crystals, and micro cavities, the refractive index can be assumed piecewise constant. In that case, the *Coulomb gauge*, $\nabla \cdot \mathbf{A}(\mathbf{r}) = 0$, is locally satisfied and only the transverse part of the electric field $\mathbf{E}_{\mathrm{T}}(\mathbf{r})$ enters in the first term of Eq. (10.68). It is convenient to express the fields in terms of vector and scalar potential $\mathbf{A}(\mathbf{r})$ and $\varphi(\mathbf{r})$ such that the two homogeneous Maxwell equations are automatically satisfied. In the generalized Coulomb gauge, the longitudinal electric field is related to $\varphi(\mathbf{r})$ which is determined by the Poisson equation and can thus be expressed solely in terms of electronic operators. This part and its contribution to the total Hamiltonian will be treated in Section 10.3.5.

The propagating part of the electric field as well as the magnetic field are completely determined by the vector potential via

$$\mathbf{E}_{\mathrm{T}} = -\frac{\partial \mathbf{A}}{\partial t}, \tag{10.69}$$

$$\mathbf{B} = -\frac{\partial \mathbf{A}}{\partial t}, \tag{10.70}$$

where the vector potential satisfies the wave equation

$$\nabla \times \nabla \times \mathbf{A}(\mathbf{r}, t) + \frac{n^2(\mathbf{r})}{c^2} \frac{\partial^2}{\partial t^2} \mathbf{A}(\mathbf{r}, t) = \mu_0 \mathbf{j}_{\mathrm{T}}. \tag{10.71}$$

Here, the speed of light in vacuum is locally modified inside different materials by the refractive index $n(\mathbf{r})$. This part is not actively interacting with the electromagnetic field such that $n(\mathbf{r})$ can be taken as independent of the energetic and temporal characteristics of the light. The coupling to the active semiconductor material is provided by the transversal current density $\mathbf{j}_{\mathrm{T}}$.

In order to obtain a suitable starting point for field quantization, we first study the free-field case in the absence of carriers, i.e., by studying the passive

dielectric structure alone. In that case, the current density $\mathbf{j}_T$ of Eq. (10.71) vanishes and the total vector potential can be expanded in the eigenfunctions of the Fourier transform of Eq. (10.71). Alternatively, the steady-state solutions to Maxwell's equations can be found via an ansatz $\mathbf{U}_{\mathbf{q}\sigma}\mathrm{e}^{-\omega_{\mathbf{q}}t}$. Inserting this into Eq. (10.71), we obtain within each layer of constant index of refraction the Helmholtz equation

$$\nabla \times \nabla \times \mathbf{U}_{\mathbf{q}\sigma}(\mathbf{r}) - q^2 n^2(\mathbf{r})\mathbf{U}_{\mathbf{q}\sigma}(\mathbf{r}) = \mathbf{0}, \tag{10.72}$$

where the three-dimensional wave vector $\mathbf{q}$ of the light mode is connected to its frequency via the relation $\omega_{\mathbf{q}} = c|\mathbf{q}|$ and σ denotes the polarization direction of the field. In the following analysis, we mostly investigate planar quantum-well structures where $n(\mathbf{r})$ is spatially varying only in z-direction.

Since Eq. (10.72) forms a generalized eigenvalue problem, the solutions form a complete set of transversal eigenfunctions which can be orthonormalized via

$$\int \mathrm{d}^3r\, n^2(z)\, \mathbf{U}^*_{\mathbf{q}\sigma}(\mathbf{r}) \cdot \mathbf{U}_{\mathbf{q}'\sigma'}(\mathbf{r}) = \delta_{\mathbf{q},\mathbf{q}'}\delta_{\sigma,\sigma'}. \tag{10.73}$$

By multiplying Eq. (10.72) by $\mathbf{U}_{\mathbf{q}'\sigma'}(\mathbf{r})$, integrating over all space, and subtracting the same term with $\mathbf{q}$ and $\mathbf{q}'$ exchanged, one can even show that

$$(|\mathbf{q}|^2 - |\mathbf{q}'|^2) \int \mathrm{d}^3r\, n^2(z)\, \mathbf{U}_{\mathbf{q}\sigma}(\mathbf{r}) \cdot \mathbf{U}_{\mathbf{q}'\sigma'}(\mathbf{r}) = 0\,, \tag{10.74}$$

which provides a generalized orthogonality relation between $\mathbf{U}_{\mathbf{q}\sigma}$ and $\mathbf{U}_{\mathbf{q}'\sigma'}$. The fact that the integral must vanish for $q \neq q'$ will be needed later on in this section.

The corresponding completeness relation from the solutions of Eq. (10.72) is given by

$$\left[\sum_{\mathbf{q},\sigma} \mathbf{U}^*_{\mathbf{q}\sigma,n}(\mathbf{r})\;\; \mathbf{U}_{\mathbf{q}'\sigma',m}(\mathbf{r}')\right] = \frac{\delta^{\mathrm{T}}_{n,m}(\mathbf{r}-\mathbf{r}')}{n^2(z)}, \tag{10.75}$$

where $\delta^{\mathrm{T}}_{n,m}(\mathbf{r})$ is the transversal δ-function [3] and the indices n and m label the x-, y-, or z-component of a three-dimensional vector. For divergence-free vector fields, it acts as a regular δ-function, and for a general field, it additionally projects onto the transverse part.

Once the eigenmodes are known, the vector potential can be expanded in terms of $\mathbf{U}_{\mathbf{q}\sigma}$. In a classical description, the mode expansion is

$$\mathbf{A}(\mathbf{r}) = \sum_{\mathbf{q}} \left[\mathbf{U}_{\mathbf{q}}(\mathbf{r})C_{\mathbf{q}}(t) + \mathbf{U}^*_{\mathbf{q}}(\mathbf{r})C^*_{\mathbf{q}}(t)\right], \tag{10.76}$$

with complex-valued coefficients $C_{\mathbf{q}(\mathbf{t})}$. To simplify the expressions, we have omitted the polarization index σ since it is often obvious which polarization direction of the light is studied. If this is not the case, one has to assume that the polarization index is implicitly included in $\mathbf{q}$. We will frequently use this implicit notation in the following derivations.

In classical optics, the coefficients $C_{\mathbf{q}}(t)$ of Eq. (10.76) are time dependent and have a precise phase and amplitude. However, the Heisenberg uncertainty principle dictates that in the quantum case each mode $C_{\mathbf{q}}(t)$ must have an uncertainty in phase and amplitude. To incorporate this intrinsic feature, we will later replace $C_{\mathbf{q}}(t)$ by an operator. In order to motivate this step, we express the classical field energy, Eq. (10.68), in terms of the time-dependent coefficients. To that aim, it is important to note that for the non-interacting case, the time evolution is known; since the functions $\mathbf{U}_{\mathbf{q}}$ are solutions to Eq. (10.72), the time evolution of the coefficients is given by a simple harmonic evolution $C_{\mathbf{q}}(t) = C_{\mathbf{q},0} \exp(-\mathrm{i}\omega_q t)$. Thus, the transverse electric field and the magnetic field which have to be inserted into Eq. (10.68) are given by

$$\mathbf{E}_{\mathrm{T}}(\mathbf{r}, t) = \mathrm{i} \sum_{\mathbf{q}} \omega_q \Big[\mathbf{U}_{\mathbf{q}}(\mathbf{r}) C_{\mathbf{q}}(t) - \mathbf{U}^*_{\mathbf{q}}(\mathbf{r}) C^*_{\mathbf{q}}(t) \Big], \tag{10.77}$$

$$\mathbf{B}(\mathbf{r}, t) = \sum_{\mathbf{q}} \omega_q \Big[\nabla \times \mathbf{U}_{\mathbf{q}}(\mathbf{r}) C_{\mathbf{q}}(t) + \nabla \times \mathbf{U}^*_{\mathbf{q}}(\mathbf{r}) C^*_{\mathbf{q}}(t) \Big]. \tag{10.78}$$

When we now insert these expressions into the electromagnetic field energy, we are careful not to switch the order of coefficients in our derivations, as these coefficients will become operators later on. The relevant integrals which need to be solved are

$$\begin{aligned}
H_{\mathrm{elec}} &= \frac{\varepsilon_0}{2} \int \mathrm{d}^3 r\, n^2(\mathbf{r}) \left(\sum_{\mathbf{q}} \omega_q \mathbf{U}_{\mathbf{q}}(\mathbf{r}) C_{\mathbf{q}} - \text{c.c.} \right) \left(\sum_{\mathbf{q}'} \omega_{q'} \mathbf{U}^*_{\mathbf{q}'}(\mathbf{r}) C^*_{\mathbf{q}'} - \text{c.c.} \right) \\
&= \frac{\varepsilon_0}{2} \sum_{\mathbf{q},\mathbf{q}'} \omega_q\, \omega_{q'} \Big\{ \left(\int \mathrm{d}^3 r\, n^2(\mathbf{r}) \mathbf{U}_{\mathbf{q}}(\mathbf{r}) \cdot \mathbf{U}^*_{\mathbf{q}'}(\mathbf{r}) \right) C_{\mathbf{q}} C^*_{\mathbf{q}'} \\
&\quad + \left(\int \mathrm{d}^3 r\, n^2(\mathbf{r}) \mathbf{U}^*_{\mathbf{q}}(\mathbf{r}) \cdot \mathbf{U}_{\mathbf{q}'}(\mathbf{r}) \right) C^*_{\mathbf{q}} C_{\mathbf{q}'} \\
&\quad - \left[\left(\int \mathrm{d}^3 r\, n^2(\mathbf{r}) \mathbf{U}_{\mathbf{q}}(\mathbf{r}) \cdot \mathbf{U}_{\mathbf{q}'}(\mathbf{r}) \right) C_{\mathbf{q}} C_{\mathbf{q}'} + \text{c.c.} \right] \Big\} \\
&= \frac{\varepsilon_0}{2} \sum_{\mathbf{q}} \omega_q^2 \Big[C_{\mathbf{q}} C^*_{\mathbf{q}} + C^*_{\mathbf{q}} C_{\mathbf{q}} \Big] \\
&\quad - \frac{\varepsilon_0}{2} \sum_{\mathbf{q},\mathbf{q}'} \omega_q \omega_{q'} \left[\left(\int \mathrm{d}^3 r\, n^2(\mathbf{r}) \mathbf{U}_{\mathbf{q}}(\mathbf{r}) \cdot \mathbf{U}_{\mathbf{q}'}(\mathbf{r}) \right) C_{\mathbf{q}} C_{\mathbf{q}'} + \text{c.c.} \right]
\end{aligned} \tag{10.79}$$

$$
\begin{aligned}
H_{\mathrm{magn}} &= \frac{\varepsilon_0}{2}\int \mathrm{d}^3r\ c^2\left(\sum_{\mathbf{q}}\ \nabla\times\mathbf{U}_{\mathbf{q}}(\mathbf{r})C_{\mathbf{q}}+\mathrm{c.c.}\right)\left(\sum_{\mathbf{q}'}\ \nabla\times\mathbf{U}^*_{\mathbf{q}'}(\mathbf{r})C^*_{\mathbf{q}'}+\mathrm{c.c.}\right)\\
&= \frac{\varepsilon_0}{2}\sum_{\mathbf{q},\mathbf{q}'}\left\{\left[\int \mathrm{d}^3r\ c^2\mathbf{U}_{\mathbf{q}}(\mathbf{r})\cdot\left(\nabla\times\nabla\times\mathbf{U}^*_{\mathbf{q}'}(\mathbf{r})\right)\right]C_{\mathbf{q}}C^*_{\mathbf{q}'}\right.\\
&\quad+\left[\int \mathrm{d}^3rc^2\mathbf{U}^*_{\mathbf{q}}(\mathbf{r})\cdot(\nabla\times\nabla\times\mathbf{U}_{\mathbf{q}'}(\mathbf{r}))\right]C^*_{\mathbf{q}}C_{\mathbf{q}'}\\
&\quad\left.+\left[\left(\int \mathrm{d}^3rc^2\mathbf{U}_{\mathbf{q}}(\mathbf{r})\cdot(\nabla\times\nabla\times\mathbf{U}_{\mathbf{q}'}(\mathbf{r}))\right)C_{\mathbf{q}}C_{\mathbf{q}'}+\mathrm{c.c.}\right]\right\}\\
&= \frac{\varepsilon_0}{2}\sum_{\mathbf{q},\mathbf{q}'}\ \omega^2_{q'}\left\{\left[\int \mathrm{d}^3rn^2(\mathbf{r})\mathbf{U}_{\mathbf{q}}(\mathbf{r})\cdot\mathbf{U}^*_{\mathbf{q}'}(\mathbf{r})\right]C_{\mathbf{q}}\ C^*_{\mathbf{q}'}\right.\\
&\quad+\left[\int \mathrm{d}^3rn^2(\mathbf{r})\mathbf{U}^*_{\mathbf{q}}(\mathbf{r})\cdot\mathbf{U}_{\mathbf{q}'}(\mathbf{r})\right]C^*_{\mathbf{q}}C_{\mathbf{q}'}\\
&\quad\left.+\left[\left(\int \mathrm{d}^3rn^2(\mathbf{r})\mathbf{U}_{\mathbf{q}}(\mathbf{r})\cdot\mathbf{U}_{\mathbf{q}'}(\mathbf{r})\right)C_{\mathbf{q}}\ C_{\mathbf{q}'}+\mathrm{c.c.}\right]\right\}\\
&= \frac{\varepsilon_0}{2}\sum_{\mathbf{q}}\omega^2_q\left[C_{\mathbf{q}}C^*_{\mathbf{q}}+C^*_{\mathbf{q}}C_{\mathbf{q}}\right]\\
&\quad+\frac{\varepsilon_0}{2}\sum_{\mathbf{q},\mathbf{q}'}\ \omega^2_{q'}\left[\left(\int \mathrm{d}^3rn^2(\mathbf{r})\mathbf{U}_{\mathbf{q}}(\mathbf{r})\cdot\mathbf{U}_{\mathbf{q}'}(\mathbf{r})\right)C_{\mathbf{q}}C_{\mathbf{q}'}+\mathrm{c.c.}\right]
\end{aligned}
\tag{10.80}
$$

If we now add up the two contributions from Eqs. (10.79) and (10.80) to calculate the total field energy, the terms in the respective last lines cancel since we have shown that the integrals only contribute for $\omega_q = \omega_{q'}$, which makes the two terms identical except for their sign.

If we furthermore introduce the dimensionless coefficients $\tilde{C}_{\mathbf{q}} = \omega_q/\mathsf{E}_q C_{\mathbf{q}}$ with the vacuum field amplitude

$$
\mathsf{E}_q = \sqrt{\frac{\hbar\omega_q}{2\varepsilon_0}},
\tag{10.81}
$$

the total field energy,

$$
H_{\mathrm{em}} = \sum_{\mathbf{q}}\frac{\hbar\omega_q}{2}\left[\tilde{C}_{\mathbf{q}}\tilde{C}^*_{\mathbf{q}}+\tilde{C}^*_{\mathbf{q}}\tilde{C}_{\mathbf{q}}\right],
\tag{10.82}
$$

exactly resembles that of an ensemble of uncoupled harmonic oscillators, where each field mode corresponds to one oscillator state.

Since each non-interacting mode behaves like a harmonic oscillator, we can now quantize the transverse electromagnetic field by introducing photon

creation and annihilation operators $\hat{B}^{\dagger}_{\mathbf{q}}$ and $\hat{B}_{\mathbf{q}}$ corresponding to the classical coefficients $\tilde{C}^{*}_{\mathbf{q}}$ and $\tilde{C}_{\mathbf{q}}$. We require that these operators satisfy the canonical Bosonic commutation relations,

$$\left[\hat{B}_{\mathbf{q}}, \hat{B}^{\dagger}_{\mathbf{q}'}\right] = \delta_{\mathbf{q},\mathbf{q}'}\delta_{\sigma,\sigma'} \tag{10.83}$$

and

$$\left[\hat{B}_{\mathbf{q}}, \hat{B}_{\mathbf{q}'}\right] = \left[\hat{B}^{\dagger}_{\mathbf{q}}, \hat{B}^{\dagger}_{\mathbf{q}'}\right] = 0. \tag{10.84}$$

For notational simplicity, we leave out the hat symbol for photon operators in the remainder of this article.

From Eq. (10.82), we can immediately read off the quantized form of the field Hamiltonian as

$$\hat{H}_{\mathrm{em}} = \sum_{\mathbf{q}} \frac{\hbar\omega_{\mathbf{q}}}{2}\left[B_{\mathbf{q}}B^{\dagger}_{\mathbf{q}} + B^{\dagger}_{\mathbf{q}}B_{\mathbf{q}}\right] = \sum_{\mathbf{q}} \hbar\omega_{\mathbf{q}}\left[B^{\dagger}_{\mathbf{q}}B_{\mathbf{q}} + \frac{1}{2}\right], \tag{10.85}$$

where we have already used the commutation relation, Eq. (10.83), once. Since the expansion coefficients are operators, also the vector potential and the electric and magnetic fields become operators. Before we introduce the final expressions, we note that for a sufficiently large in-plane extension $\mathsf{S} = \mathsf{L}^2$, the eigenmode solutions can be separated into in-plane and z-dependent parts:

$$\mathbf{U}_{\mathbf{q}\sigma}(\mathbf{r}_{\|}, z) = \frac{1}{\sqrt{\mathsf{S}}}\, \mathrm{e}^{\mathrm{i}\mathbf{q}_{\|}\cdot\mathbf{r}_{\|}}\, \mathbf{u}_{\mathbf{q}\sigma}(z)\,, \tag{10.86}$$

where $\mathbf{q} = (\mathbf{q}_{\|}, q_z)$. The remaining z-dependent component can be computed for example with the help of the transfer-matrix technique outlined, e.g., in Kira et al. [7]. With these mode functions, the operator expansions for the vector potential with explicit inclusion of the polarization index are given by

$$\hat{\mathbf{A}}(\mathbf{r}, z) = \sum_{\mathbf{q}_{\|},q_z,\sigma} \frac{\mathsf{E}_{\mathbf{q}}}{\omega_{\mathbf{q}}}\left[\mathbf{u}_{\mathbf{q},\sigma}(z)\frac{\mathrm{e}^{\mathrm{i}\mathbf{q}_{\|}\cdot\mathbf{r}_{\|}}}{\sqrt{\mathsf{S}}} B_{\mathbf{q}_{\|},q_z,\sigma} + \mathbf{u}^{*}_{\mathbf{q},\sigma}(z)\frac{\mathrm{e}^{-\mathrm{i}\mathbf{q}_{\|}\cdot\mathbf{r}_{\|}}}{\sqrt{\mathsf{S}}} B^{\dagger}_{\mathbf{q}_{\|},q_z\sigma}\right]. \tag{10.87}$$

The magnetic field is more easily expressed as

$$\begin{aligned} \hat{\mathbf{B}}(\mathbf{r}) &= \nabla \times \hat{\mathbf{A}} \\ &= \sum_{\mathbf{q}_{\|},q_z,\sigma} \frac{\mathsf{E}_{\mathbf{q}}}{\omega_{\mathbf{q}}}\left[\left(\nabla \times \mathbf{U}_{\mathbf{q},\sigma}(\mathbf{r})\right)B_{\mathbf{q}_{\|},q_z,\sigma} + \left(\nabla \times \mathbf{U}^{*}_{\mathbf{q},\sigma}(\mathbf{r})\right)B^{\dagger}_{\mathbf{q}_{\|},q_z,\sigma}\right] \end{aligned} \tag{10.88}$$

in terms of the full mode functions $\mathbf{U}_{\mathbf{q},\sigma}$. In practice, it is rarely needed.

Since the electric field operator involves a time derivative, we have to know the Heisenberg equations of motion for the newly defined photon operators. Using the field Hamiltonian, Eq. (10.85), and the commutation relations, Eqs. (10.83) and (10.84), we easily obtain the operator dynamics

$$\mathrm{i}\hbar \frac{\partial}{\partial t} B_{\mathbf{q}_{\|},q_z} = \hbar\omega_{\mathbf{q}} B_{\mathbf{q}_{\|},q_z} \tag{10.89}$$

and

$$\mathrm{i}\hbar \frac{\partial}{\partial t} B^{\dagger}_{\mathbf{q}_{\|},q_z} = -\hbar\omega_{\mathbf{q}} B^{\dagger}_{\mathbf{q}_{\|},q_z}, \tag{10.90}$$

and from here the final expression for the electric field

$$\begin{aligned} \hat{\mathbf{E}}_{\mathrm{T}}(\mathbf{r}) &= -\frac{\partial}{\partial t}\hat{\mathbf{A}}(\mathbf{r}) = \frac{\mathrm{i}}{\hbar}\left[\hat{\mathbf{A}}(\mathbf{r}), \hat{H}_{\mathrm{em}}\right] \\ &= \sum_{\mathbf{q}_{\|},q_z,\sigma} \mathrm{i}\mathsf{E}_{\mathbf{q}} \left(\mathbf{u}_{\mathbf{q},\sigma}(z)\frac{\mathrm{e}^{\mathrm{i}\mathbf{q}_{\|}\cdot\mathbf{r}_{\|}}}{\sqrt{\mathsf{S}}} B_{\mathbf{q}_{\|},q_z,\sigma} - \mathbf{u}^{*}_{\mathbf{q},\sigma}(z)\frac{\mathrm{e}^{-\mathrm{i}\mathbf{q}_{\|}\cdot\mathbf{r}_{\|}}}{\sqrt{\mathsf{S}}} B^{\dagger}_{\mathbf{q}_{\|},q_z\sigma}\right). \end{aligned} \tag{10.91}$$

The field Hamiltonian in terms of the operators $\hat{\mathbf{E}}_{\mathrm{T}}$ and $\hat{\mathbf{B}}$ has exactly the same form as Eq. (10.68), but with the classical fields replaced by the corresponding operators, i.e.,

$$\hat{H}_{\mathrm{em}} = \frac{\varepsilon_0}{2}\int_{\mathsf{L}^3} \left[n^2(z)\hat{\mathbf{E}}^2_{\mathrm{T}}(\mathbf{r},t) + c^2\hat{\mathbf{B}}^2(\mathbf{r},t)\right] \mathrm{d}^3 r. \tag{10.92}$$

This Hamiltonian yields the correct energy contribution of the photon field alone also in the case of an interacting system. The actual interaction Hamiltonian between light and semiconductor electrons is discussed in Section 10.3.2.

For later reference, we note at this point that the quantized light field is always truly three dimensional even when one studies dimensionally reduced semiconductor systems, such as quantum wells, wires, or dots. The special system geometry enters only into the optical part of the description when the mode functions are computed from Eq. (10.72).

10.2.8 Second Quantization of Lattice Vibrations

Even though the periodic lattice of atoms can often be assumed to be perfect, the ions still oscillate around their equilibrium positions. Once the position of an ion is disturbed from its equilibrium value, it is pulled back via the collective Coulomb interaction with the rest of the ions. As long as the displacement is not too large, the equilibrating force can be approximated as a harmonic force

leading to collective vibrations in the solid. Since the optically active electrons can interact with these lattice vibrations, we need to include them explicitly in many calculations where we want to realistically analyze experimentally relevant situations.

For this purpose, we now abandon the mean-field description of the ions for a while and consider their microscopic treatment for solids, first for the simplest case where we have one atom inside the unit cell. The results for more than one atom per unit cell will be mentioned at the end of this section. As the simplest model of ionic motion, we assume that ions at different lattice sites are coupled harmonically. In first quantization, the Hamiltonian of N ions has the form

$$\hat{H}_{\mathrm{ph}} = \sum_j \frac{\hat{\mathbf{P}}_j^2}{2M} + \frac{1}{2}\sum_{n\neq m}\frac{1}{2}M\Omega_{n-m}^2(\Delta\hat{\mathbf{R}}_n - \Delta\hat{\mathbf{R}}_m)^2, \tag{10.93}$$

where $\mathbf{P}_j$ is the momentum of ion j with mass M. In this Hamiltonian, the ion at position $\mathbf{R}_j$ deviates from its equilibrium value $\mathbf{R}_j^0$ by the distance $\Delta\mathbf{R}_j = \mathbf{R}_j - \mathbf{R}_j^0$. The two-particle interaction introduces a harmonic force with respect to the deviations $\Delta\mathbf{R}_j$. Here, we assume that the harmonic term depends only on the relative distance between the lattice sites such that the coupling has the form Ω_{n-m}^2.

The lattice vibrations are quantized by introducing the usual canonical commutation relations,

$$\left[\Delta\hat{\mathbf{R}}_{n,\alpha}, \hat{\mathbf{P}}_{m,\beta}\right] = \mathrm{i}\hbar\delta_{n,m}\,\delta_{\alpha,\beta} \tag{10.94}$$

and

$$\left[\Delta\hat{\mathbf{R}}_{n,\alpha}, \Delta\hat{\mathbf{R}}_{m,\beta}\right] = \left[\hat{\mathbf{P}}_{n,\alpha}, \hat{\mathbf{P}}_{m,\beta}\right] = 0, \tag{10.95}$$

where α and β refer to the usual Cartesian components x, y, and z. Since Eq. (10.93) represents a genuine many-body system and the two-particle interaction is harmonic, it is – once again – convenient to adopt the formalism of second quantization. For this purpose, we introduce an annihilation operator

$$D_{\mathbf{p},\sigma} = \frac{-\mathrm{i}}{\sqrt{N}}\sum_{j=1}^{N} \mathrm{e}^{-\mathrm{i}\mathbf{R}_j^0\cdot\mathbf{p}}\left(\sqrt{\frac{M\Omega_{\mathbf{p}}}{2\hbar}}\Delta\hat{\mathbf{R}}_j + \mathrm{i}\frac{\hat{\mathbf{P}}_j}{\sqrt{2\,\hbar\Omega_{\mathbf{p}}M}}\right)\cdot\mathbf{e}_{\mathbf{p},\sigma} \tag{10.96}$$

and a creation operator

$$D_{\mathbf{p},\sigma}^{\dagger} = \frac{\mathrm{i}}{\sqrt{N}}\sum_{j=1}^{N} \mathrm{e}^{\mathrm{i}\mathbf{R}_j^0\cdot\mathbf{p}}\left(\sqrt{\frac{M\Omega_{\mathbf{p}}}{2\hbar}}\Delta\hat{\mathbf{R}}_j - \mathrm{i}\frac{\hat{\mathbf{P}}_j}{2\sqrt{\hbar\Omega_{\mathbf{p}}M}}\right)\cdot\mathbf{e}_{\mathbf{p},\sigma}, \tag{10.97}$$

for phonons, where $\mathbf{e}_{\mathbf{p},\sigma}$ defines the directions of the vibration identified by σ. These definitions are nothing but a many-body generalization of the usual second quantization of a single harmonic oscillator. We observe that D and $D^{\dagger}$ involve all lattice sites such that these operators represent collective vibrations. The associated quasi-particles are the *phonons*. The quantity $\Omega_{\mathbf{p}}$ is the collective phonon frequency with the property $\Omega_{\mathbf{p}} = \Omega_{-\mathbf{p}}$; this is proven later when $\Omega_{\mathbf{p}}$ is computed explicitly.

Next, we check the commutation relations between the phonon operators:

$$
\begin{aligned}
\left[D_{\mathbf{p}}, D^{\dagger}_{\mathbf{p}'}\right] =& \frac{1}{N}\sum_{n,m} \mathrm{e}^{-\mathrm{i}\mathbf{R}^0_n\cdot\mathbf{p}+\mathrm{i}\mathbf{R}^0_m\cdot\mathbf{p}'}\mathbf{e}_{\mathbf{p}}\cdot\left(-\frac{\mathrm{i}}{2\hbar}\sqrt{\frac{\Omega_{\mathbf{p}}}{\Omega_{\mathbf{p}'}}}\left[\Delta\hat{\mathbf{R}}_n, \hat{\mathbf{P}}_m\right]\right. \\
&\left.+\frac{\mathrm{i}}{2\hbar}\sqrt{\frac{\Omega_{\mathbf{p}'}}{\Omega_{\mathbf{p}}}}\left[\hat{\mathbf{P}}_n, \Delta\hat{\mathbf{R}}_m\right]\right)\cdot\mathbf{e}_{\mathbf{p}'} \\
=& \frac{1}{N}\sum_{n,m} \mathrm{e}^{-\mathrm{i}\mathbf{R}^0_n\cdot\mathbf{p}+\mathrm{i}\mathbf{R}^0_m\cdot\mathbf{p}'}\left(-\frac{\mathrm{i}}{2\hbar}\sqrt{\frac{\Omega_{\mathbf{p}}}{\Omega_{\mathbf{p}'}}}\,\mathrm{i}\hbar\delta_{n,m}\right. \\
&\left.+\frac{\mathrm{i}}{2\hbar}\sqrt{\frac{\Omega_{\mathbf{p}'}}{\Omega_{\mathbf{p}}}}\left(-\mathrm{i}\hbar\delta_{n,m}\right)\right)\mathbf{e}_{\mathbf{p}}\cdot\mathbf{e}_{\mathbf{p}'} \\
=& \frac{1}{N}\sum_{n} \mathrm{e}^{-\mathrm{i}\mathbf{R}^0_n\cdot(\mathbf{p}-\mathbf{p}')}\left(\frac{1}{2}\sqrt{\frac{\Omega_{\mathbf{p}}}{\Omega_{\mathbf{p}'}}}+\frac{1}{2}\sqrt{\frac{\Omega_{\mathbf{p}'}}{\Omega_{\mathbf{p}}}}\right)\mathbf{e}_{\mathbf{p}}\cdot\mathbf{e}_{\mathbf{p}'} \\
=& \delta_{\mathbf{p},\mathbf{p}'}\left(\frac{1}{2}\sqrt{\frac{\Omega_{\mathbf{p}}}{\Omega_{\mathbf{p}'}}}+\frac{1}{2}\sqrt{\frac{\Omega_{\mathbf{p}'}}{\Omega_{\mathbf{p}}}}\right)\mathbf{e}_{\mathbf{p}}\cdot\mathbf{e}_{\mathbf{p}} = \delta_{\mathbf{p},\mathbf{p}'},
\end{aligned}
\tag{10.98}
$$

where we have used the definitions (10.96) and (10.97) together with the commutation relations (10.94) and (10.95). Once again, we have introduced the implicit notation where the phonon branch index σ is included in $\mathbf{p}$. A similar derivation yields

$$
\left[D_{\mathbf{p}}, D_{\mathbf{p}'}\right] = \left[D^{\dagger}_{\mathbf{p}}, D^{\dagger}_{\mathbf{p}'}\right] = 0. \tag{10.99}
$$

Thus, the phonon operators $D_{\mathbf{p}}$ and $D^{\dagger}_{\mathbf{p}}$ obey bosonic commutation relations, which was expected for the harmonic interaction potential.

Our next task is to express the Hamiltonian (10.93) in terms of phonon operators. In order to do this, we have to express the individual $\Delta\hat{\mathbf{R}}_n$ and $\hat{\mathbf{P}}_n$ via phonon operators. For this purpose, we consider the completeness relation for plane waves on a periodic lattice,

$$
\frac{1}{N}\sum_{\mathbf{p}} \mathrm{e}^{\mathrm{i}\left(\mathbf{R}^0_n-\mathbf{R}^0_m\right)\cdot\mathbf{p}} = \delta_{n,\,m}. \tag{10.100}
$$

With help of this, we express the displacement and momentum operators by using

$$\Delta\hat{\mathbf{R}}_n = \frac{\mathrm{i}}{\sqrt{N}}\sum_{\mathbf{p}}\sqrt{\frac{\hbar}{2M\Omega_{\mathbf{p}}}}\left(D_{\mathbf{p}}\mathrm{e}^{\mathrm{i}\mathbf{R}_n^0\cdot\mathbf{p}}\mathbf{e}_{\mathbf{p}} - D^{\dagger}_{\mathbf{p}}\mathrm{e}^{-\mathrm{i}\mathbf{R}_n^0\cdot\mathbf{p}}\mathbf{e}_{\mathbf{p}}\right)$$
$$= \frac{\mathrm{i}}{\sqrt{N}}\sum_{\mathbf{p}}\sqrt{\frac{\hbar}{2M\Omega_{\mathbf{p}}}}\left(D_{\mathbf{p}}\mathbf{e}_{\mathbf{p}} - D^{\dagger}_{-\mathbf{p}}\mathbf{e}_{-\mathbf{p}}\right)\mathrm{e}^{\mathrm{i}\mathbf{R}_n^0\cdot\mathbf{p}} \tag{10.101}$$

and

$$\hat{\mathbf{P}}_n = \frac{1}{\sqrt{N}}\sum_{\mathbf{p}}\sqrt{\frac{\hbar M\Omega_{\mathbf{p}}}{2}}\left(D_{\mathbf{p}}\mathrm{e}^{\mathrm{i}\mathbf{R}_n^0\cdot\mathbf{p}}\mathbf{e}_{\mathbf{p}} + D^{\dagger}_{\mathbf{p}}\mathrm{e}^{-\mathrm{i}\mathbf{R}_n^0\cdot\mathbf{p}}\mathbf{e}_{\mathbf{p}}\right)$$
$$= \frac{1}{\sqrt{N}}\sum_{\mathbf{p}}\sqrt{\frac{\hbar M\Omega_{\mathbf{p}}}{2}}\left(D_{\mathbf{p}}\mathbf{e}_{\mathbf{p}} + D^{\dagger}_{-\mathbf{p}}\mathbf{e}_{-\mathbf{p}}\right)\mathrm{e}^{\mathrm{i}\mathbf{R}_n^0\cdot\mathbf{p}}. \tag{10.102}$$

In these derivations, we utilized the property that the sum over $\mathbf{p}$ includes both $+\mathbf{p}$ and $-\mathbf{p}$. These relations can now be inserted into Eq. (10.93), which leads to

$$\hat{H}_{\mathrm{ph}} = \sum_{\mathbf{p},\mathbf{p}'}\frac{1}{2M}\frac{\hbar M}{2}\sqrt{\Omega_{\mathbf{p}}\Omega_{\mathbf{p}'}}\left(D_{\mathbf{p}}\mathbf{e}_{\mathbf{p}} + D^{\dagger}_{-\mathbf{p}}\mathbf{e}_{-\mathbf{p}}\right)$$
$$\cdot\left(D_{-\mathbf{p}'}\mathbf{e}_{-\mathbf{p}'} + D^{\dagger}_{\mathbf{p}'}\mathbf{e}_{\mathbf{p}'}\right)\sum_{n}\frac{1}{N}\mathrm{e}^{\mathrm{i}\mathbf{R}_n^0\cdot(\mathbf{p}-\mathbf{p}')}$$
$$-\frac{M}{2}\sum_{\mathbf{p},\mathbf{p}'}\frac{\hbar}{2M}\frac{1}{\sqrt{\Omega_{\mathbf{p}}\Omega_{\mathbf{p}'}}}\left(D_{\mathbf{p}}\mathbf{e}_{\mathbf{p}} - D^{\dagger}_{-\mathbf{p}}\mathbf{e}_{-\mathbf{p}}\right)$$
$$\cdot\left(D_{-\mathbf{p}'}\mathbf{e}_{-\mathbf{p}'} - D^{\dagger}_{\mathbf{p}'}\mathbf{e}_{\mathbf{p}'}\right)$$
$$\times\sum_{n,m}\frac{1}{N}\mathrm{e}^{\mathrm{i}\mathbf{R}_n^0\cdot(\mathbf{p}-\mathbf{p}')}\frac{\Omega^2_{n-m}}{2}\left(1-\mathrm{e}^{\mathrm{i}\left(\mathbf{R}_m^0-\mathrm{i}\mathbf{R}_n^0\right)\cdot\mathbf{p}}\right)$$
$$\times\left(1-\mathrm{e}^{-\mathrm{i}\left(\mathbf{R}_m^0-\mathrm{i}\mathbf{R}_n^0\right)\cdot\mathbf{p}'}\right). \tag{10.103}$$

The sums over the lattice sites can be performed analytically. For the first sum, we find

$$\Sigma_1 = \frac{1}{N}\sum_{n}\mathrm{e}^{\mathrm{i}\mathbf{R}_n^0\cdot(\mathbf{p}-\mathbf{p}')} = \delta_{\mathbf{p},\mathbf{p}'}, \tag{10.104}$$

and the second sum yields

$$\Sigma_2 = \sum_{n,m} \frac{1}{N} \mathrm{e}^{\mathrm{i}\mathbf{R}_n^0 \cdot (\mathbf{p}-\mathbf{p}')} \frac{\Omega_{n-m}^2}{2} \left(1 - \mathrm{e}^{\mathrm{i}\left(\mathbf{R}_m^0 - \mathrm{i}\mathbf{R}_n^0\right)\cdot\mathbf{p}}\right)\left(1 - \mathrm{e}^{-\mathrm{i}\left(\mathbf{R}_m^0 - \mathrm{i}\mathbf{R}_n^0\right)\cdot\mathbf{p}'}\right)$$
$$= \sum_{n} \frac{1}{N} \mathrm{e}^{\mathrm{i}\mathbf{R}_n^0 \cdot (\mathbf{p}-\mathbf{p}')} \sum_{\Delta m} \frac{\Omega_{\Delta m}^2}{2} \left(1 - \mathrm{e}^{\mathrm{i}\left(\mathbf{R}_{\Delta m}^0\right)\cdot\mathbf{p}}\right)\left(1 - \mathrm{e}^{-\mathrm{i}\left(\mathbf{R}_{\Delta m}^0\right)\cdot\mathbf{p}'}\right)$$
$$= \delta_{\mathbf{p},\mathbf{p}'} \sum_m \frac{\Omega_m^2}{2} \left|1 - \mathrm{e}^{\mathrm{i}\mathbf{R}_m^0\cdot\mathbf{p}}\right|^2. \tag{10.105}$$

By identifying explicitly the collective phonon frequency,

$$\Omega_{\mathbf{p}} \equiv \sqrt{\sum_m \frac{\Omega_m^2}{2} \left|1 - \mathrm{e}^{\mathrm{i}\mathbf{R}_m^0\cdot\mathbf{p}}\right|^2}, \tag{10.106}$$

we can now simplify the phonon Hamiltonian into the form

$$\hat{H}_{\mathrm{ph}} = \sum_{\mathbf{p}} \frac{1}{4} \hbar\Omega_{\mathbf{p}} \left(D_{\mathbf{p}}\mathbf{e}_{\mathbf{p}} + D_{-\mathbf{p}}^\dagger \mathbf{e}_{-\mathbf{p}}\right) \cdot \left(D_{-\mathbf{p}}\mathbf{e}_{-\mathbf{p}} + D_{\mathbf{p}}^\dagger \mathbf{e}_{\mathbf{p}}\right)$$
$$- \sum_{\mathbf{p}} \frac{1}{4} \hbar\Omega_{\mathbf{p}} \left(D_{\mathbf{p}}\mathbf{e}_{\mathbf{p}} - D_{-\mathbf{p}}^\dagger \mathbf{e}_{-\mathbf{p}}\right) \cdot \left(D_{-\mathbf{p}}\mathbf{e}_{-\mathbf{p}} - D_{\mathbf{p}}^\dagger \mathbf{e}_{\mathbf{p}}\right)$$
$$= \sum_{\mathbf{p}} \frac{1}{2} \hbar\Omega_{\mathbf{p}} \left(D_{\mathbf{p}} D_{\mathbf{p}}^\dagger \;\; \mathbf{e}_{\mathbf{p}} \cdot \mathbf{e}_{\mathbf{p}} + D_{-\mathbf{p}}^\dagger D_{-\mathbf{p}} \;\; \mathbf{e}_{-\mathbf{p}} \cdot \mathbf{e}_{-\mathbf{p}}\right)$$
$$= \sum_{\mathbf{p}} \hbar\Omega_{\mathbf{p}} \left(D_{\mathbf{p}}^\dagger D_{\mathbf{p}} + \frac{1}{2}\right), \tag{10.107}$$

where we used the commutation relation (10.98). Once again, we find a Hamiltonian describing a set of harmonic oscillators.

In order to evaluate the eigenfrequencies $\Omega_{\mathbf{p}}$, we now take the long-wavelength limit of the dispersion relation (10.106). As a result, we find a linear dependency,

$$\Omega_{\mathbf{p}} = c_{\mathrm{A}} |\mathbf{p}|, \tag{10.108}$$

where the coefficient c_{A} is called the (acoustic) phonon velocity of sound. Due to this special dispersion, these lattice vibrations are called acoustic phonons. Since the velocity of sound can be experimentally measured for different materials, we use these experimental values in Eqs. (10.107) and (10.108) whenever we treat acoustic phonon effects.

Even in confined systems, the lattice vibrations propagate through the entire three-dimensional sample. Thus, the phonons in most quantum-well, quantum-wire, and embedded quantum-dot systems are truly three dimensional. Consequently, Eq. (10.107) can be used as a starting point without dimension-dependent modifications.

In many semiconductor lattices, we have unit cells consisting of more than one atom. In this case, in addition to the acoustic also optical phonon excitations exist. As it turns out, the optical phonon dispersion is nearly independent of **p** in contrast to the acoustic phonons. In GaAs-based systems, optical phonons have an energy around 36 meV.

If the lattice is close to thermal equilibrium, the phonon occupation numbers closely follow the Bose–Einstein distribution:

$$\langle D_{\mathbf{p}}^{\dagger} D_{\mathbf{p}} \rangle = \frac{1}{e^{\frac{\hbar\Omega_{\mathbf{p}}}{k_B T}} - 1}, \tag{10.109}$$

where k_B is the Boltzmann constant and T is the temperature of the sample. Hence, for low temperatures (below approximately 100 K), optical phonon populations are often negligible since their occupation $\langle D_{\mathbf{p}}^{\dagger} D_{\mathbf{p}} \rangle \ll 1$. In these cases, it is often sufficient to focus on the effects of acoustic phonons only. Since most of the quantum-optical investigations are performed under such conditions, we do not derive the optical phonon effects in detail.

10.3 Interactions in Semiconductors

In this section, we complete the derivation of the basic Hamiltonian for semiconductor quantum optics. Building on the concepts introduced in the previous section, we now focus on the interaction aspects. Hence, we not only have to formulate the light–matter coupling Hamiltonian in general, but we also have to describe the interactions in the electronic system, i.e., the carrier–carrier Coulomb and the carrier–phonon interaction.

We use the second quantization formalism for the electrons, photons, and phonons in the solid. The discussion is presented explicitly for quantum-well systems. However, once we have the explicit expressions, it is relatively straightforward to generalize them to obtain the Hamiltonians for systems with other effective dimensionalities.

10.3.1 Many-Body Hamiltonian

Starting from the Hamiltonian in first quantization, Eq. (10.4), we apply the relations (10.34) and (10.35) to write the total system Hamiltonian in second quantized notation:

$$\hat{H} = \hat{H}_{\rm em} + \hat{H}_{\rm ph} + \int \hat{\Psi}^\dagger(\mathbf{r}) \left[\frac{\hat{\mathbf{p}}^2}{2m_0} + \tilde{V}_L(\mathbf{r}) \right] \hat{\Psi}(\mathbf{r}) \mathrm{d}^3 r$$
$$+ \int \hat{\Psi}^\dagger(\mathbf{r}) \left[-\frac{Q}{m_0} \hat{\mathbf{A}}(\mathbf{r}) \cdot \hat{\mathbf{p}} + \frac{Q^2}{2m_0} \hat{\mathbf{A}}^2(\mathbf{r}) \right] \hat{\Psi}(\mathbf{r}) \mathrm{d}^3 r$$
$$+ \int \hat{\Psi}^\dagger(\mathbf{r}) \hat{\Psi}^\dagger(\mathbf{r}')\; V(\mathbf{r} - \mathbf{r}')\; \hat{\Psi}(\mathbf{r}) \hat{\Psi}(\mathbf{r}') \mathrm{d}^3 r\, \mathrm{d}^3 r', \tag{10.110}$$

where $\hat{H}_{\rm em}$ and $\hat{H}_{\rm ph}$ are given by Eqs. (10.85) and (10.107), respectively. In order to include the electron–phonon interaction, we allow $\tilde{V}_{\rm L}(\mathbf{r})$ to describe lattice vibrations in the form

$$\tilde{V}_{\rm L}(\mathbf{r}) = \sum_n U\Big(\mathbf{r} - \Big(\mathbf{R}_n^0 + \Delta \mathbf{R}_n\Big)\Big), \tag{10.111}$$

where $U(\mathbf{r})$ is the effective potential of one ion at the origin and the deviation from the equilibrium positions $\mathbf{R}_n^0$ is expressed in terms of the displacement vector $\Delta\mathbf{R}_n$. In the ground state, every ion is located at its equilibrium position, i.e., $\Delta\mathbf{R}_n \equiv 0$, and $\tilde{V}_{\rm L}(\mathbf{r})$ is equal to the original lattice periodic potential which we used to define the Bloch basis. For small deviations of the lattice ions from their equilibrium position, a Taylor expansion of the ion potential yields

$$\tilde{V}_{\rm L}(\mathbf{r}) = \sum_n U\Big(\mathbf{r} - \mathbf{R}_n^0 - \Delta\mathbf{R}_n\Big)$$
$$= \sum_n U\Big(\mathbf{r} - \mathbf{R}_n^0\Big) - \sum_n \nabla U\Big(\mathbf{r} - \mathbf{R}_n^0\Big) \cdot \Delta\mathbf{R}_n + \mathrm{O}\Big(\Delta\mathbf{R}_n^2\Big)$$
$$= V_{\rm L}(\mathbf{r}) - \sum_n \nabla U\Big(\mathbf{r} - \mathbf{R}_n^0\Big) \cdot \Delta\mathbf{R}_n + \mathrm{O}\Big(\Delta\mathbf{R}_n^2\Big). \tag{10.112}$$

where in the last step we have identified the original lattice periodic potential.

Since the lattice vibrations are quantized according to Section 10.2.8, we actually have to use the operator form $\Delta\hat{\mathbf{R}}_n$ defined by Eq. (10.101). If we neglect the higher order corrections to the Taylor expansion, the system Hamiltonian can be written in the form

$$\hat{H} = \hat{H}_{\rm em} + \hat{H}_{\rm ph} + \int \hat{\Psi}^\dagger(\mathbf{r}) \left[\frac{\hat{\mathbf{p}}^2}{2m_0} + V_{\rm L}(\mathbf{r}) \right] \hat{\Psi}(\mathbf{r}) \mathrm{d}^3 r$$
$$+ \int \hat{\Psi}^\dagger(\mathbf{r}) \left[-\frac{Q}{m_0} \hat{\mathbf{A}}(\mathbf{r}) \cdot \hat{\mathbf{p}} + \frac{Q^2}{2m_0} \hat{\mathbf{A}}^2(\mathbf{r}) \right] \hat{\Psi}(\mathbf{r}) \mathrm{d}^3 r$$
$$- \int \hat{\Psi}^\dagger(\mathbf{r}) \left[\sum_n \nabla U(\mathbf{r} - \mathbf{R}_n^0) \cdot \Delta\hat{\mathbf{R}}_n \right] \hat{\Psi}(\mathbf{r}) \mathrm{d}^3 r$$
$$+ \int \hat{\Psi}^\dagger(\mathbf{r}) \hat{\Psi}^\dagger(\mathbf{r}')\; V(\mathbf{r} - \mathbf{r}')\; \hat{\Psi}(\mathbf{r}) \hat{\Psi}(\mathbf{r}') \mathrm{d}^3 r\, \mathrm{d}^3 r'. \tag{10.113}$$

With this arrangement, the first line contains the Hamiltonians of the non-interacting systems, the second line is the light–electron interaction, the third line describes electron–phonon interaction, while the last line contains the Coulomb interaction among the active electrons in their different bands. Each of these interaction terms will be evaluated explicitly in the following sections.

10.3.2 Light–Matter Interaction

The coupling of electrons and light is described by the second line of Eq. (10.113). In order to formulate this interaction Hamiltonian more explicitly, we have to look at both contributions:

$$\hat{H}_{\mathrm{A\cdot p}} = -\int \hat{\Psi}^{\dagger}(\mathbf{r})\left[\frac{Q}{m_0}\hat{\mathbf{A}}(\mathbf{r})\cdot\hat{\mathbf{p}}\right]\hat{\Psi}(\mathbf{r})\mathrm{d}^3r \tag{10.114}$$

and

$$\hat{H}_{\mathrm{A\cdot A}} = \int \hat{\Psi}^{\dagger}(\mathbf{r})\left[\frac{Q^2}{2m_0}\hat{\mathbf{A}}^2(\mathbf{r})\right]\hat{\Psi}(\mathbf{r})\mathrm{d}^3r. \tag{10.115}$$

Since we want to choose quantum wells as our representative semiconductor system, we use the explicit Bloch-electron wavefunction, Eq. (10.39).

With this choice, we can write Eqs. (10.114) and (10.115) in the generic form:

$$\begin{aligned}\hat{H}_j &= \int \hat{\Psi}^{\dagger}(\mathbf{r})O_j(\mathbf{r})\hat{\Psi}(\mathbf{r})\mathrm{d}^3r \\ &= \sum_{\lambda,\mathbf{k}_\|,\lambda',\mathbf{k}'_\|} a^{\dagger}_{\lambda,\mathbf{k}_\|}a_{\lambda',\mathbf{k}'_\|} I^{\lambda,\mathbf{k}_\|}_{\lambda',\mathbf{k}'_\|}\big|_j,\end{aligned} \tag{10.116}$$

with the matrix element between Bloch electrons

$$I^{\lambda,\mathbf{k}_\|}_{\lambda',\mathbf{k}'_\|}\big|_j \equiv \int \varphi^{*}_{\lambda,\mathbf{k}_\|}(\mathbf{r})O_j(\mathbf{r})\varphi_{\lambda',\mathbf{k}'_\|}(\mathbf{r})\mathrm{d}^3r. \tag{10.117}$$

In the following, we will analyze this integral for $O_j(\mathbf{r}) = -\frac{Q}{m_0}\hat{\mathbf{A}}(\mathbf{r})\cdot\hat{\mathbf{p}}$ and $O_j(\mathbf{r}) = \frac{Q^2}{2m_0}\hat{\mathbf{A}}^2(\mathbf{r})$. For this purpose, we first introduce a Fourier decomposition of the vector potential in order to separate the different $\mathbf{q}_\|$ contributions:

$$\hat{\mathbf{A}}(\mathbf{r},z) = \sum_{\mathbf{q}_\|} \hat{\mathbf{A}}_{\mathbf{q}_\|}(z)\mathrm{e}^{\mathrm{i}\mathbf{q}_\|\cdot\mathbf{r}_\|}. \tag{10.118}$$

With the help of Eq. (10.87), we see that

$$\hat{\mathbf{A}}_{\mathbf{q}_\|}(z) = \sum_{q_z}\frac{1}{\sqrt{\mathsf{S}}}\frac{\mathsf{E}_{\mathbf{q}}}{\omega_{\mathbf{q}}}\left[\mathbf{u}_{\mathbf{q}}(z)B_{\mathbf{q}_\|,q_z} + \mathbf{u}^{*}_{-\mathbf{q}_\|,q_z}(z)B^{\dagger}_{-\mathbf{q}_\|,q_z}\right]. \tag{10.119}$$

Using this form in Eq. (10.117) together with the explicit envelope function (10.39), we obtain

$$I^{\lambda,\mathbf{k}_\parallel}_{\lambda',\mathbf{k}'_\parallel}|_{\mathrm{A\cdot p}} = -\frac{Q}{m_0}\sum_{\mathbf{q}_\parallel}\int \frac{\mathrm{e}^{-\mathrm{i}\mathbf{k}_\parallel\cdot\mathbf{r}_\parallel}}{\sqrt{\mathsf{S}}}\xi^*(z)u^*_{\lambda,\mathbf{k}_\parallel}(\mathbf{r})\mathrm{e}^{\mathrm{i}\mathbf{q}_\parallel\cdot\mathbf{r}_\parallel}$$
$$\times \quad \hat{\mathbf{A}}_{\mathbf{q}_\parallel}(z)\cdot\hat{\mathbf{p}}\varphi_{\lambda',\mathbf{k}'_\parallel}(\mathbf{r})\mathrm{d}^3r. \tag{10.120}$$

Next, we have to evaluate $\hat{\mathbf{p}}$ acting on $\varphi_{\lambda',\mathbf{k}'_\parallel}(\mathbf{r})$. By separating the different parts, $\hat{\mathbf{p}} = \hat{\mathbf{p}}_\parallel - \mathrm{i}\hbar\mathbf{e}_z\frac{\partial}{\partial z}$, we find

$$\begin{aligned}
\sqrt{\mathsf{S}}\hat{\mathbf{p}}\left[\varphi_{\lambda',\mathbf{k}'_\parallel}(\mathbf{r})\right] &= \hat{\mathbf{p}}\left[\mathrm{e}^{\mathrm{i}\mathbf{k}'_\parallel\cdot\mathbf{r}_\parallel}\xi_{\lambda'}(z)u_{\lambda',\mathbf{k}'_\parallel}(\mathbf{r})\right] \\
&= \left(\hat{\mathbf{p}}_\parallel - \mathrm{i}\hbar\mathbf{e}_z\frac{\partial}{\partial z}\right)\left[\mathrm{e}^{\mathrm{i}\mathbf{k}'_\parallel\cdot\mathbf{r}_\parallel}\xi_{\lambda'}(z)u_{\lambda',\mathbf{k}'_\parallel}(\mathbf{r})\right] \\
&= \mathrm{e}^{\mathrm{i}\mathbf{k}'_\parallel\cdot\mathbf{r}_\parallel}\xi_{\lambda'}(z)\left[\hbar\mathbf{k}'_\parallel + \hat{\mathbf{p}}\right]u_{\lambda',\mathbf{k}'_\parallel}(\mathbf{r}) \\
&\quad - \mathrm{i}\hbar\mathbf{e}_z\mathrm{e}^{\mathrm{i}\mathbf{k}'_\parallel\cdot\mathbf{r}_\parallel}u_{\lambda',\mathbf{k}'_\parallel}(\mathbf{r})\frac{\partial}{\partial z}\xi_{\lambda'}(z)\,.
\end{aligned} \tag{10.121}$$

To identify the different parts, we write Eq. (10.120) as

$$I^{\lambda,\mathbf{k}_\parallel}_{\lambda',\mathbf{k}'_\parallel}|_{\mathrm{A\cdot p}} \equiv I^{\lambda,\mathbf{k}_\parallel}_{\lambda',\mathbf{k}'_\parallel}|_{\mathrm{A\cdot p(1)}} + I^{\lambda,\mathbf{k}_\parallel}_{\lambda',\mathbf{k}'_\parallel}|_{\mathrm{A\cdot p(2)}}, \tag{10.122}$$

where

$$I^{\lambda,\mathbf{k}_\parallel}_{\lambda',\mathbf{k}'_\parallel}|_{\mathrm{A\cdot p(1)}} = -\frac{Q}{m_0}\frac{1}{\mathsf{S}}\sum_{\mathbf{q}_\parallel}\int \mathrm{e}_\parallel^{\mathrm{i}(\mathbf{k}'_\parallel+\mathbf{q}_\parallel-\mathbf{k}_\parallel)\cdot\mathbf{r}_\parallel}\xi^*_\lambda(z)\hat{\mathbf{A}}_{\mathbf{q}_\parallel}(z)\xi_{\lambda'}(z)\cdot u^*_{\lambda,\mathbf{k}_\parallel}(\mathbf{r})$$
$$\times \quad \left[\hbar\mathbf{k}'_\parallel + \mathbf{p}\right]u_{\lambda',\mathbf{k}'_\parallel}(\mathbf{r})\mathrm{d}^3r \tag{10.123}$$

and

$$I^{\lambda,\mathbf{k}_\parallel}_{\lambda',\mathbf{k}'_\parallel}|_{\mathrm{A\cdot p(2)}} = \frac{\mathrm{i}\hbar Q}{m_0}\frac{1}{\mathsf{S}}\sum_{\mathbf{q}_\parallel}\int \mathrm{e}^{\mathrm{i}(\mathbf{k}'_\parallel+\mathbf{q}_\parallel-\mathbf{k}_\parallel)\cdot\mathbf{r}_\parallel}\xi^*_\lambda(z)$$
$$\times \quad \hat{\mathbf{A}}_{\mathbf{q}_\parallel}(z)\cdot\mathbf{e}_z\frac{\partial\xi_{\lambda'}(z)}{\partial z}u^*_{\lambda,\mathbf{k}_\parallel}(\mathbf{r})u_{\lambda',\mathbf{k}'_\parallel}(\mathbf{r})\mathrm{d}^3r. \tag{10.124}$$

Clearly, these two expressions correspond to the different parts of $\hat{\mathbf{p}}$ in Eq. (10.121).

In order to complete the light–matter Hamiltonian, we still have to express Eq. (10.115) in the Bloch basis. This procedure introduces a matrix element

$$I^{\lambda,\mathbf{k}_\parallel}_{\lambda',\mathbf{k}'_\parallel}|_{\mathrm{A\cdot A}} = \frac{1}{\mathsf{S}}\sum_{\mathbf{q}_\parallel,\mathbf{q}'_\parallel}\int \mathrm{e}^{\mathrm{i}(\mathbf{k}'_\parallel+\mathbf{q}_\parallel-\mathbf{q}'_\parallel-\mathbf{k}_\parallel)\cdot\mathbf{r}_\parallel}\,\frac{Q^2}{2m_0}\hat{\mathbf{A}}_{\mathbf{q}_\parallel}(z)\cdot\hat{\mathbf{A}}_{-\mathbf{q}'_\parallel}(z)$$
$$\times \quad \xi^*_\lambda(z)u^*_{\lambda,\mathbf{k}_\parallel}(\mathbf{r})\xi_{\lambda'}(z)u_{\lambda',\mathbf{k}'_\parallel}\mathrm{d}^3r \tag{10.125}$$

similar to $I|_{\mathrm{A\cdot p}}$.

On our way to solve the remaining integrals (10.123), (10.124), and (10.125), let us consider the generic form

$$I^{\lambda,\mathbf{k}_\parallel}_{\lambda',\mathbf{k}'_\parallel}|_G = \frac{1}{\mathsf{S}}\int \mathrm{e}^{\mathrm{i}\Delta\mathbf{Q}_\parallel\cdot\mathbf{r}_\parallel}\,\left[\xi^*_\lambda(z)\hat{A}(z)\xi_{\lambda'}(z)\right]$$
$$\times\left[u^*_{\lambda,\mathbf{k}_\parallel}(\mathbf{r})C(\mathbf{r})u_{\lambda',\mathbf{k}'_\parallel}(\mathbf{r})\right]\mathrm{d}^3r. \tag{10.126}$$

Here, the different terms vary on quite different length scales: The quantity $A(z)$ changes on the scale of the light field, i.e., the optical wavelength; the lattice periodic Bloch functions $u_{\lambda,\mathbf{k}_\parallel}$ and $C(\mathbf{r})$ vary on the scale of the atomic unit cell; and the factor $\mathrm{e}^{\mathrm{i}\Delta\mathbf{Q}_\parallel\cdot\mathbf{r}_\parallel}$ varies on the mesoscopic scale of the envelope function and the plane-wave part of the light field.

We will now make use of these different characteristic length scales to simplify the overall integration. For this purpose, we divide the integral (10.126) into parts over each unit-cell volume $v_\mathbf{R}$ centered at lattice point $\mathbf{R} = (\mathbf{R}_\parallel, Z)$:

$$I^{\lambda,\,\mathbf{k}_\parallel}_{\lambda',\,\mathbf{k}'_\parallel}|_G = \frac{1}{\mathsf{S}}\sum_{\mathbf{R}}\int_{v_\mathbf{R}} \mathrm{e}^{\mathrm{i}\Delta\mathbf{Q}_\parallel\cdot\mathbf{r}_\parallel}\,\left[\xi^*_\lambda(z)\hat{A}(z)\xi_{\lambda'}(z)\right]$$
$$\times\left[u^*_{\lambda,\,\mathbf{k}_\parallel}(\mathbf{r})C(\mathbf{r})u_{\lambda',\,\mathbf{k}'_\parallel}(\mathbf{r})\right]\mathrm{d}^3r. \tag{10.127}$$

Since only $u_{\lambda,\mathbf{k}_\parallel}$ and $C(\mathbf{r})$ vary within a unit cell, the remaining terms can be taken to be constant over the $v_\mathbf{R}$ integration such that

$$I^{\lambda,\,\mathbf{k}_\parallel}_{\lambda',\,\mathbf{k}'_\parallel}|_G = \frac{1}{\mathsf{S}}\sum_{\mathbf{R}} \mathrm{e}^{\mathrm{i}\Delta\mathbf{Q}_\parallel\cdot\mathbf{R}_\parallel}\,\left[\xi^*_\lambda(Z)\hat{A}(Z)\xi_{\lambda'}(Z)\right]v_\mathbf{R}$$
$$\times\frac{1}{v_\mathbf{R}}\int_{v_\mathbf{R}} u^*_{\lambda,\,\mathbf{k}_\parallel}(\mathbf{r})C(\mathbf{r})u_{\lambda',\,\mathbf{k}'_\parallel}\mathrm{d}^3r. \tag{10.128}$$

Since $u_{\lambda,\mathbf{k}_\parallel}$ and $C(\mathbf{r})$ are lattice periodic, the $v_\mathbf{R}$ integrals are equal for all lattice sites. Hence, we may introduce the position-independent matrix element:

$$\langle\lambda,\mathbf{k}_\parallel|\hat{C}|\lambda',\mathbf{k}'_\parallel\rangle \equiv \frac{1}{v_0}\int_{v_0} u^*_{\lambda,\,\mathbf{k}_\parallel}(\mathbf{r})C(\mathbf{r})u_{\lambda',\,\mathbf{k}'_\parallel}\mathrm{d}^3r, \tag{10.129}$$

where v_0 denotes the common unit-cell volume. This volume is infinitesimal compared to the remaining terms in (10.128), so that we can use $v_0 = \mathrm{d}^2\mathbf{R}_\parallel \mathrm{d}Z$ and convert the sum into an integral. With these modifications, we find

$$I^{\lambda,\,\mathbf{k}_\parallel}_{\lambda',\,\mathbf{k}'_\parallel}|_G = \left(\frac{1}{\mathsf{S}}\int \mathrm{e}^{\mathrm{i}\Delta\mathbf{Q}_\parallel\cdot\mathbf{R}_\parallel}\mathrm{d}^2 R_\parallel\right) A^{\lambda,\lambda'}\ \langle\lambda,\mathbf{k}_\parallel|\hat{C}|\lambda',\mathbf{k}'_\parallel\rangle, \tag{10.130}$$

where we have identified the envelope-function matrix element of $\hat{A}(z)$ via

$$A^{\lambda,\lambda'} \equiv \int \xi^*_\lambda(z) A(z) \xi_{\lambda'}(z)\mathrm{d}z, \tag{10.131}$$

which only depends on the confinement structure. Furthermore, the $\mathbf{R}_\parallel$ integration can be evaluated analytically by noting that

$$\frac{1}{\mathsf{S}}\int \mathrm{e}^{\mathrm{i}\Delta\mathbf{Q}_\parallel\cdot\mathbf{R}_\parallel}\mathrm{d}^2 R_\parallel = \delta_{\Delta\mathbf{Q}_\parallel,0}. \tag{10.132}$$

With the help of these relations, the integral (10.130) becomes

$$I^{\lambda,\,\mathbf{k}_\parallel}_{\lambda',\,\mathbf{k}'_\parallel}|_G(\mathbf{q}) = \delta_{\Delta\mathbf{Q}_\parallel,0}\ A^{\lambda,\lambda'}\ \langle\lambda,\mathbf{k}_\parallel|\hat{C}|\lambda',\mathbf{k}'_\parallel\rangle. \tag{10.133}$$

This result can be used directly to generate the explicit forms of $I^{\lambda,\,\mathbf{k}_\parallel}_{\lambda',\,\mathbf{k}'_\parallel}|_G$ once $\mathbf{Q}_\parallel$, $\hat{A}(z)$, and $C(\mathbf{r})$ are identified.

For $I_{\mathrm{A\cdot p(1)}}$, the symbolic factor $\mathrm{e}^{\mathrm{i}\Delta\mathbf{Q}_\parallel\cdot\mathbf{r}_\parallel}$ stands for

$$\sum_{\mathbf{q}_\parallel} \mathrm{e}^{\mathrm{i}(\mathbf{q}_\parallel-\mathbf{k}_\parallel+\mathbf{k}'_\parallel)\cdot\mathbf{r}_\parallel}, \tag{10.134}$$

while

$$\hat{A}(z) = -\frac{Q}{m_0}\hat{\mathbf{A}}_{\mathbf{q}_\parallel}(z), \quad \hat{C} = \hbar\mathbf{k}'_\parallel + \hat{\mathbf{p}}. \tag{10.135}$$

In the same way, we identify

$$\sum_{\mathbf{q}_\parallel} \mathrm{e}^{\mathrm{i}(\mathbf{q}_\parallel-\mathbf{k}_\parallel+\mathbf{k}'_\parallel)\cdot\mathbf{r}_\parallel}, \quad \hat{A}(z) = \frac{\mathrm{i}\hbar Q}{m_0}\hat{\mathbf{A}}_{\mathbf{q}_\parallel}(z)\cdot\mathbf{e}_z\frac{\partial}{\partial z}, \quad \hat{C} = 1 \tag{10.136}$$

for the matrix element $I_{\mathrm{A\cdot p(2)}}$ and

$$\sum_{\mathbf{q}_\parallel,\mathbf{q}'_\parallel} \mathrm{e}^{\mathrm{i}(\mathbf{q}_\parallel-\mathbf{q}'_\parallel-\mathbf{k}_\parallel+\mathbf{k}'_\parallel)\cdot\mathbf{r}_\parallel}, \quad \hat{A}(z) = \frac{Q^2}{2m_0}\hat{\mathbf{A}}_{\mathbf{q}_\parallel}(z)\cdot\hat{\mathbf{A}}_{-\mathbf{q}'_\parallel}(z), \quad \hat{C} = 1 \tag{10.137}$$

for the matrix element $I_{\mathrm{A\cdot A}}$. Thus, we obtain the explicit results:

$$I^{\lambda,\,\mathbf{k}_\|}_{\lambda',\,\mathbf{k}'_\|}\big|_{\mathrm{A\cdot p(1)}} = -\sum_{\mathbf{q}_\|}\delta_{\mathbf{k}'_\|,\mathbf{k}_\|-\mathbf{q}_\|}\frac{Q}{m_0} \times \hat{\mathbf{A}}^{\lambda,\lambda'}_{\mathbf{q}_\|}\cdot\langle\lambda,\mathbf{k}_\||\Big[\hbar(\mathbf{k}_\|-\mathbf{q}_\|)+\hat{\mathbf{p}}\Big]|\mathbf{k}_\|-\mathbf{q}_\|,\lambda'\rangle, \tag{10.138}$$

$$I^{\lambda,\mathbf{k}_\|}_{\lambda',\mathbf{k}'_\|}\big|_{\mathrm{A\cdot p(2)}} = \sum_{\mathbf{q}_\|}\delta_{\mathbf{k}'_\|,\mathbf{k}_\|-\mathbf{q}_\|}\frac{\mathrm{i}\hbar Q}{m_0}\int \xi^*_\lambda(z)\hat{\mathbf{A}}_{\mathbf{q}_\|}(z)\cdot\mathbf{e}_z\frac{\partial\xi_{\lambda'}(z)}{\partial z}\;\mathrm{d}z \times\langle\lambda,\mathbf{k}_\||\mathbf{k}_\|-\mathbf{q}_\|,\lambda'\rangle, \tag{10.139}$$

$$I^{\lambda,\mathbf{k}_\|}_{\lambda',\mathbf{k}'_\|}\big|_{\mathrm{A\cdot A}} = \sum_{\mathbf{q}_\|,\mathbf{q}'_\|}\delta_{\mathbf{k}'_\|-\mathbf{q}'_\|,\mathbf{k}_\|-\mathbf{q}_\|}\frac{Q^2}{2m_0}A^{(2),\lambda,\lambda'}_{\mathbf{q}_\|,-\mathbf{q}'_\|} \times\langle\lambda,\mathbf{k}_\||\mathbf{k}_\|+\mathbf{q}'_\|-\mathbf{q}_\|,\lambda'\rangle. \tag{10.140}$$

According to Eq. (10.131), the different confinement matrix elements of the vector potential are given by

$$\hat{\mathbf{A}}^{\lambda,\lambda'}_{\mathbf{q}_\|} = \int \xi^*_\lambda(z)\hat{\mathbf{A}}_{\mathbf{q}_\|}(z)\xi_{\lambda'}(z)\mathrm{d}z \equiv A^{\lambda,\lambda'}_{\mathbf{q}_\|}\mathbf{e}_\mathrm{P}, \tag{10.141}$$

$$\hat{A}^{(2),\lambda,\lambda'}_{\mathbf{q}_\|,-\mathbf{q}'_\|} = \int \xi^*_\lambda(z)\hat{\mathbf{A}}_{\mathbf{q}_\|}(z)\cdot\hat{\mathbf{A}}_{-\mathbf{q}'_\|}(z)\xi_{\lambda'}(z)\mathrm{d}z, \tag{10.142}$$

which defines the polarization direction $\mathbf{e}_\mathrm{P}$ of the field.

The final form of the matrix elements (10.138), (10.139), and (10.140) can be computed once the specific forms of the Bloch functions are known. For this, we need information about the band structure, which we use at the level of the $\mathbf{k}\cdot\mathbf{p}$ results as discussed in Section 10.2.3. Before we evaluate Eqs. (10.138), (10.139), and (10.140), we note that in all matrix elements over the unit cell the index of the Bloch functions differs only by the parallel momentum of the light field which is roughly two orders of magnitude smaller than typical carrier momenta. Hence, it is a good approximation to evaluate

$$\langle\lambda,\mathbf{k}_\||\hat{C}|\mathbf{k}'_\|,\lambda'\rangle = \langle\lambda,\mathbf{k}_\||\hat{C}|\mathbf{k}_\|-\mathbf{q}_\|,\lambda'\rangle \approx \langle\lambda,\mathbf{k}_\||\hat{C}|\mathbf{k}_{\|,\lambda'}\rangle\,. \tag{10.143}$$

More rigorously, this can be justified by a Taylor expansion of the more symmetric form $\langle\lambda,\mathbf{k}_\|+\frac{\mathbf{q}_\|}{2}|\hat{C}|\mathbf{k}_\|-\frac{\mathbf{q}_\|}{2},\lambda'\rangle$ of the matrix element.

With the help of the $(\mathbf{k}\cdot\mathbf{p})$-function (10.18), we obtain

$$
\begin{aligned}
\langle\lambda, \mathbf{k}_\parallel|[\hbar\mathbf{k} + \hat{\mathbf{p}}]|\mathbf{k}_\parallel, \lambda'\rangle &= \hbar\mathbf{k}_\parallel\langle\lambda, \mathbf{k}_\parallel|\mathbf{k}_\parallel, \lambda'\rangle + \langle\lambda, \mathbf{k}_\parallel|\hat{\mathbf{p}}|\mathbf{k}_\parallel, \lambda'\rangle \\
&= \hbar\mathbf{k}_\parallel\delta_{\lambda,\lambda'} + \langle\lambda|\hat{\mathbf{p}}|\lambda'\rangle \\
&\quad + \frac{\hbar}{m_0}\left(\sum_{\eta\neq\lambda}\frac{\langle\lambda|\hat{\mathbf{p}}|\eta\rangle\cdot\mathbf{k}_\parallel\langle\eta|\hat{\mathbf{p}}|\lambda'\rangle}{\epsilon_0^\lambda - \epsilon_0^\eta} + \sum_{\eta\neq\lambda'}\frac{\langle\lambda|\hat{\mathbf{p}}|\eta\rangle\mathbf{k}_\parallel\cdot\langle\eta|\hat{\mathbf{p}}|\lambda'\rangle}{\epsilon_0^{\lambda'} - \epsilon_0^\eta}\right) \\
&\quad + \mathsf{O}(\mathbf{k}^2) \\
&= \hbar\mathbf{k}_\parallel\delta_{\lambda,\lambda'} + \langle\lambda|\hat{\mathbf{p}}|\lambda'\rangle \\
&\quad + \frac{\hbar\mathbf{k}_\parallel}{m_0}\left(\sum_{\eta\neq\lambda}\frac{\langle\lambda|\hat{p}|\eta\rangle\langle\eta|\hat{p}|\lambda'\rangle}{\epsilon_0^\lambda - \epsilon_0^\eta} + \sum_{\eta\neq\lambda'}\frac{\langle\lambda|\hat{p}|\eta\rangle\langle\eta|\hat{p}|\lambda'\rangle}{\epsilon_0^{\lambda'} - \epsilon_0^\eta}\right) + \mathsf{O}(\mathbf{k}^2)\,,
\end{aligned}
\tag{10.144}
$$

where the last form follows if the isotropic approximation, Eq. (10.29), is made. The parity of the Bloch functions (10.17) implies that $\langle\lambda|\hat{\mathbf{p}}|\lambda'\rangle$ vanishes for equal band indices. Thus, it is convenient to separate the $\lambda = \lambda'$ and $\lambda \neq \lambda'$ parts. This procedure leads to

$$
\begin{aligned}
\langle\lambda, \mathbf{k}_\parallel|[\hbar\mathbf{k} + \hat{\mathbf{p}}]|\mathbf{k}_\parallel, \lambda'\rangle &= \delta_{\lambda,\lambda'}\hbar\mathbf{k}_\parallel\left[1 + \frac{2}{m_0}\sum_{\eta\neq\lambda}\frac{\langle\lambda|\hat{p}|\eta\rangle\langle\eta|\hat{p}|\lambda\rangle}{\epsilon_0^\lambda - \epsilon_0^\eta}\right] \\
&\quad + \left(1 - \delta_{\lambda,\lambda'}\right)\langle\lambda|\hat{\mathbf{p}}|\lambda'\rangle + \mathsf{O}(\mathbf{k}^2),
\end{aligned}
\tag{10.145}
$$

where we have restricted the analysis to a two-band model. We can easily convince ourselves that in that case the last line of Eq. (10.144) does not contribute to the term with $\lambda \neq \lambda'$. The term in square brackets in the first line on the right hand side can be expressed using the effective mass (10.27). Hence, we are left with the simple expression

$$
\langle\lambda, \mathbf{k}_\parallel|[\hbar\mathbf{k} + \hat{\mathbf{p}}]|\mathbf{k}_\parallel, \lambda'\rangle = \delta_{\lambda,\lambda'}\hbar\mathbf{k}_\parallel\frac{m_0}{m_\lambda} + \left(1 - \delta_{\lambda,\lambda'}\right)\mathbf{p}_{\lambda,\lambda'}, \tag{10.146}
$$

where we introduced the momentum matrix element

$$
\mathbf{p}_{\lambda,\lambda'} \equiv \langle\lambda|\hat{\mathbf{p}}|\lambda'\rangle. \tag{10.147}
$$

Using Eq. (10.143) as well as the explicit expressions (10.7) and (10.146), we are finally able to evaluate the different integrals in Eqs. (10.138), (10.139), and (10.140):

$$
\begin{aligned}
I_{\lambda',\mathbf{k}'_\parallel}^{\lambda,\mathbf{k}_\parallel}\big|_{\mathrm{A\cdot p(1)}} = -\sum_{\mathbf{q}_\parallel}\delta_{\mathbf{k}'_\parallel,\mathbf{k}_\parallel-\mathbf{q}_\parallel}\Bigg[&\delta_{\lambda,\lambda'}Q\frac{\hbar\mathbf{k}_\parallel}{m_\lambda}\cdot\hat{\mathbf{A}}_{\mathbf{q}_\parallel}^{\lambda,\lambda} \\
&+(1-\delta_{\lambda,\lambda'})\frac{Q\mathbf{p}_{\lambda,\lambda'}}{m_0}\cdot\hat{\mathbf{A}}_{\mathbf{q}_\parallel}^{\lambda,\lambda'}\Bigg],
\end{aligned}
\tag{10.148}
$$

$$I_{\lambda',\mathbf{k}'_{\|}}^{\lambda,\mathbf{k}_{\|}}|_{\mathrm{A\cdot p(2)}} = \frac{\mathrm{i}\hbar Q}{m_0}\delta_{\mathbf{k}'_{\|},\mathbf{k}_{\|}-\mathbf{q}_{\|}}\ \delta_{\lambda,\lambda'}\int \xi_\lambda^*(z)\hat{\mathbf{A}}_{\mathbf{q}_{\|}}(z)\cdot\mathbf{e}_z\frac{\partial\xi_\lambda(z)}{\partial z}\mathrm{d}z, \tag{10.149}$$

$$I_{\lambda',\mathbf{k}'_{\|}}^{\lambda,\mathbf{k}_{\|}}|_{\mathrm{A\cdot A}} = \sum_{\mathbf{q}_{\|},\mathbf{q}'_{\|}}\delta_{\mathbf{k}'_{\|}-\mathbf{q}'_{\|},\mathbf{k}_{\|}-\mathbf{q}_{\|}}\delta_{\lambda,\lambda'}\frac{Q^2}{2\,m_0}A^{(2),\lambda,\lambda}_{\mathbf{q}_{\|},-\mathbf{q}'_{\|}}. \tag{10.150}$$

Since $\hat{\mathbf{A}}_{\mathbf{q}_{\|}}(z)$ varies slowly on the mesoscopic scale, the last unknown integration in Eq. (10.149) can be simplified via

$$\int \xi_\lambda^*(z)\hat{\mathbf{A}}_{\mathbf{q}_{\|}}(z)\cdot\mathbf{e}_z\frac{\partial\xi_\lambda(z)}{\partial z}\mathrm{d}z = \hat{\mathbf{A}}_{\mathbf{q}_{\|}}(z_{\mathrm{QW}})\cdot\mathbf{e}_z\int \xi_\lambda^*(z)\frac{\partial\xi_\lambda(z)}{\partial z}\mathrm{d}z, \tag{10.151}$$

where z_{QW} denotes the position of the center of the quantum well. Furthermore, the confinement wave functions can be chosen real, which yields the additional simplification:

$$\begin{aligned}\hat{\mathbf{A}}_{\mathbf{q}_{\|}}(z_{\mathrm{QW}})\cdot\mathbf{e}_z\int_{-\infty}^{+\infty}\xi_\lambda^*(z)\frac{\partial\xi_\lambda(z)}{\partial z}\mathrm{d}z &= \hat{\mathbf{A}}_{\mathbf{q}_{\|}}(z_{\mathrm{QW}})\cdot\mathbf{e}_z\int_{-\infty}^{+\infty}\frac{1}{2}\frac{\partial}{\partial z}|\xi_\lambda(z)|^2\mathrm{d}z\\ &= \hat{\mathbf{A}}_{\mathbf{q}_{\|}}(z_{\mathrm{QW}})\cdot\mathbf{e}_z\Big|_{\infty}^{+\infty}|\xi_\lambda(z)|^2 = 0,\end{aligned} \tag{10.152}$$

i.e., this expression vanishes because the confinement wave function decays to zero for large distances. As a result, $I_{\mathrm{A\cdot p(2)}}$ vanishes such that

$$\begin{aligned}I_{\lambda',\mathbf{k}'_{\|}}^{\lambda,\mathbf{k}_{\|}}|_{\mathrm{A\cdot p}} &= I_{\lambda',\mathbf{k}'_{\|}}^{\lambda,\mathbf{k}_{\|}}|_{\mathrm{A\cdot p(1)}}\\ &= -\sum_{\mathbf{q}_{\|}}\delta_{\mathbf{k}'_{\|},\mathbf{k}_{\|}-\mathbf{q}_{\|}}\Bigg[\delta_{\lambda,\lambda'}Q\frac{\hbar\mathbf{k}_{\|}}{m_\lambda}\cdot\hat{\mathbf{A}}^{\lambda,\lambda}_{\mathbf{q}_{\|}}\\ &\quad + (1-\delta_{\lambda,\lambda'})\frac{Q\mathbf{p}_{\lambda,\lambda'}}{m_0}\cdot\hat{\mathbf{A}}^{\lambda,\lambda'}_{\mathbf{q}_{\|}}\Bigg].\end{aligned} \tag{10.153}$$

Collecting all the results obtained for the matrix elements, we are now able to construct the final form of the light–matter interaction Hamiltonian. This Hamiltonian follows from Eqs. (10.114), (10.115), and (10.116), where we insert the matrix elements (10.150) and (10.153), leading to

$$\begin{aligned}\hat{H}_{\mathrm{em-e}} &= \sum_{\lambda,\mathbf{k}_{\|}}\sum_{\lambda',\mathbf{k}'_{\|}}\Big[I_{\lambda',\mathbf{k}'_{\|}}^{\lambda,\mathbf{k}_{\|}}|_{\mathrm{A\cdot p}} + I_{\lambda',\mathbf{k}'_{\|}}^{\lambda,\mathbf{k}_{\|}}|_{\mathrm{A\cdot A}}\Big]a^\dagger_{\lambda,\mathbf{k}_{\|}}a_{\lambda',\mathbf{k}'_{\|}}\\ &= -\sum_{\mathbf{q}_{\|},\mathbf{k}_{\|}}\sum_{\lambda}Q\frac{\hbar\mathbf{k}_{\|}}{m_\lambda}\cdot\mathbf{e}_{\mathrm{P}}\ \hat{A}^{\lambda,\lambda}_{\mathbf{q}_{\|}}\ \ a^\dagger_{\lambda,\mathbf{k}_{\|}+\frac{\mathbf{q}_{\|}}{2}}a_{\lambda,\mathbf{k}_{\|}-\frac{\mathbf{q}_{\|}}{2}}\end{aligned}$$

$$-\sum_{\mathbf{q}_{\|},\mathbf{k}_{\|}}\sum_{\lambda\neq\lambda'}\frac{Q\mathbf{p}_{\lambda,\lambda'}}{m_0}\cdot\hat{\mathbf{A}}^{\lambda,\lambda'}_{\mathbf{q}_{\|}}\;a^{\dagger}_{\lambda,\mathbf{k}_{\|}}a_{\lambda',\mathbf{k}_{\|}-\mathbf{q}_{\|}}$$

$$+\sum_{\mathbf{q}_{\|},\mathbf{q}'_{\|}}\sum_{\lambda,\mathbf{k}_{\|}}\frac{Q^2}{2m_0}\hat{\mathrm{A}}^{(2),\lambda,\lambda}_{\mathbf{q}_{\|},-\mathbf{q}'_{\|}}\;a^{\dagger}_{\lambda,\mathbf{k}_{\|}+\mathbf{q}_{\|}}a_{\lambda,\mathbf{k}_{\|}+\mathbf{q}'_{\|}}. \tag{10.154}$$

Here, we used Eq. (10.141) to identify the polarization direction $\mathbf{e}_{\mathrm{P}}$ of the field.

In the first term of Eq. (10.154), $A^{\lambda,\lambda}_{\mathbf{q}_{\|}}$ involves either annihilation of a photon with in-plane momentum $\mathbf{q}_{\|}$ or creation of a photon with momentum $-\mathbf{q}_{\|}$. In both cases, the overall momentum conservation is assured by the corresponding changes in carrier momenta. In other words, an electron within a single band makes a transition from a state $\mathbf{k}_{\|}-\mathbf{q}_{\|}/2$ to a state $\mathbf{k}_{\|}+\mathbf{q}_{\|}/2$. This process has an *intraband* character with $\lambda=\lambda'$, and the electron momentum is changed by $\mathbf{q}_{\|}$. This process is analogous to the depicted phonon emission process on the left hand side of Fig. 10.2. These intraband transitions are proportional to the current-matrix element

$$j_{\lambda}(\mathbf{k}_{\|})\equiv\frac{Q\hbar\mathbf{k}_{\|}}{m_{\lambda}}\cdot\mathbf{e}_{\mathrm{P}}, \tag{10.155}$$

which contains the effective mass of the electron in the band λ.

The other intraband transitions follow from the last line of Eq. (10.154). In this contribution, $\hat{\mathbf{A}}^{(2)}$ contains two-photon processes where the total in-plane momentum is changed by $\mathbf{q}_{\|}-\mathbf{q}'_{\|}$. Once again, the total in-plane momentum is conserved due to the momentum exchange of electrons in the band λ. As a distinct feature of the $\hat{\mathbf{A}}^2$ interaction, we notice that the carrier part involves the free-electron mass m_0 in contrast to the $j\hat{A}$ interaction.

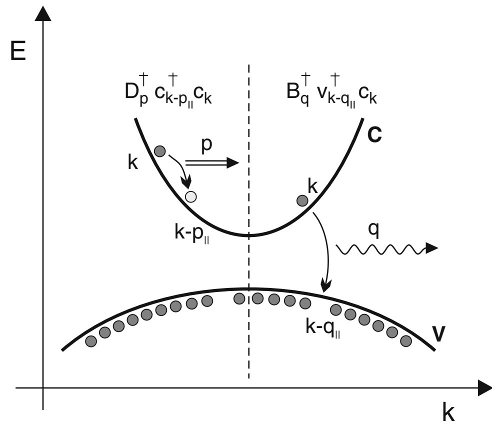


Fig. 10.2 Schematic sketch of semiconductor band structure with typical interaction terms from the light–matter (*right*) and phonon (*left*) interaction Hamiltonians. While an electron changes band and/or momentum $\hbar\mathbf{k}$, it transfers its parallel momentum to the emitted quasi-particle

The remaining parts of the light–matter Hamiltonian describe processes where photon emission or absorption is accompanied by electronic transitions between two different bands, i.e., these processes have an *interband* character. Once again, only combinations where the in-plane momentum is conserved are allowed, as depicted in the right hand part of Fig. 10.2.

For quantum-well systems without disorder, the conservation of the in-plane momentum is a general feature of the light–matter interaction. This conservation law reflects the fact that a quantum well has translational symmetry in the *xy*-plane. Due to this feature, we are able to express the in-plane parts of the electron envelope function and light modes as plane waves, which introduces a well-defined in-plane momentum for both of these entities. Since the ight–matter interaction does not break the translational symmetry, the total momentum of any allowed process must conserve the total in-plane momentum. However, due to the confinement, the quantum well does not have translational symmetry in *z*-direction, such that the *z*-component of the momentum is not conserved.

For lower dimensional systems, the momentum conservation becomes even more incomplete. For quantum wires, the momentum is conserved only along the *z*-axis parallel to the wire. In quantum dots, the translational symmetry is completely lost such that the photon momentum is entirely disconnected from the carrier system. In the other extreme, i.e., in three-dimensional bulk semiconductors, one has a complete translational symmetry in all directions such that the momentum conservation requirement has to be fulfilled for the full three-dimensional momentum vector.

As a common feature in all dimensions, the energy of the photon has to roughly match the energy difference of the carrier states participating in either intra- or interband transitions. This sets the basic energy scales of the different processes.

In our semiconductor quantum-optical investigations, we use the light–matter interaction in the form

$$\begin{aligned}\hat{H}_{\mathrm{em-e}} = \sum_{\lambda,\mathbf{k}_{\|}} \Bigg[& -\sum_{\mathbf{q}_{\|}} j_{\lambda}(\mathbf{k}_{\|})\, \hat{A}^{\lambda,\lambda}_{\mathbf{q}_{\|}}\; a^{\dagger}_{\lambda,\mathbf{k}_{\|}+\frac{\mathbf{q}_{\|}}{2}} a_{\lambda,\mathbf{k}_{\|}-\frac{\mathbf{q}_{\|}}{2}} \\ & + \sum_{\mathbf{q}_{\|},\mathbf{q}'_{\|}} \frac{Q^2}{2m_0} \hat{A}^{(2),\lambda,\lambda}_{\mathbf{q}_{\|},-\mathbf{q}'_{\|}}\; \hat{A} a^{\dagger}_{\lambda,\mathbf{k}_{\|}+\mathbf{q}_{\|}} a_{\lambda,\mathbf{k}_{\|}+\mathbf{q}'_{\|}} \Bigg] \\ & - \sum_{\mathbf{q}_{\|},\mathbf{k}_{\|}} \sum_{\lambda \neq \lambda'} \frac{Q\mathbf{p}_{\lambda,\lambda'}}{m_0} \cdot \hat{A}^{\lambda,\lambda'}_{\mathbf{q}_{\|}}\; a^{\dagger}_{\lambda,\mathbf{k}_{\|}} a_{\lambda',\mathbf{k}_{\|}-\mathbf{q}_{\|}}, \end{aligned} \tag{10.156}$$

which is organized such that the terms in the bracket describe the intraband and the other term describes the interband transitions. Since interband transitions change the energy of the carrier system roughly by the value of the band gap

$|\epsilon_0^{\lambda} - \epsilon_0^{\lambda'}|$, the electromagnetic field coupled to such a transition must be nearly resonant with the gap energy. For direct semiconductors, the typical range of this energy is roughly one up to a few electron volts which corresponds to infrared to visible and even near ultraviolet light. Thus, interband transitions can be observed and generated with optical light fields in the petahertz $(0.1 - 1 \times 10^{15}\ \mathrm{s}^{-1})$ frequency range.

The intraband transitions of carriers have a significantly lower energy, typically in the 1–100 meV range such that the corresponding photons are in the teraherz (THz) regime with a frequency range $(0.1 - 10 \times 10^{12}\ \mathrm{s}^{-1})$. Since the optical and THz field are energetically well separated, they lead to very different excitation and emission dynamics.

10.3.3 Electric Dipole Interaction

Oftentimes, the starting point for classical or quantum-optical investigations is given by the light–matter interaction in the alternative form

$$\hat{H}_D = -\int \hat{\Psi}^{\dagger}(\mathbf{r})[-e\mathbf{r} \cdot \mathbf{E}(\mathbf{r})]\hat{\Psi}(\mathbf{r})\mathrm{d}^3 r\,, \tag{10.157}$$

of the conventional dipole interaction. For classical fields, this form can be obtained from the original $(\mathbf{p} \cdot \mathbf{A})$-interaction by the so-called Goeppert-Mayer gauge transformation [3, 8]. For fields which are resonant with interband transitions, the dominant contribution involves the dipole matrix element

$$\mathbf{d}_{\lambda,\lambda'} = (-e)\langle\lambda|\mathbf{r}|\lambda'\rangle \tag{10.158}$$

for unequal $\lambda \neq \lambda'$. We can derive a similar form from Eq. (10.156) directly and without gauge transformation by noting that

$$\begin{aligned}
\frac{Q\mathbf{p}_{\lambda,\lambda'}}{m_0} &= \frac{Q}{m_0}\langle\lambda|\hat{\mathbf{p}}|\lambda'\rangle \\
&= \frac{Q}{m_0}\left\langle\lambda\left|\frac{m_0}{\mathrm{i}\hbar}\left[\hat{\mathbf{r}}, \frac{\hat{\mathbf{p}}\cdot\hat{\mathbf{p}}}{2m_0} + \hat{V}_{\mathrm{L}}(\hat{\mathbf{r}})\right]\right|\lambda'\right\rangle \\
&= \frac{Q}{\mathrm{i}\hbar}\langle\lambda|\hat{\mathbf{r}}(E_{\lambda'} - E_{\lambda})|\lambda'\rangle \\
&= -\mathrm{i}\frac{E_{\lambda'} - E_{\lambda}}{\hbar}\langle\lambda|Q\hat{\mathbf{r}}|\lambda'\rangle = \mathrm{i}\omega_{\lambda,\lambda'}\mathbf{d}_{\lambda,\lambda'}\,,
\end{aligned} \tag{10.159}$$

where we have identified the energy difference $\omega_{\lambda,\lambda'} = (E_\lambda - E_{\lambda'})/\hbar$. Now, the interband contribution of Eq. (10.156) can also be written as

$$\hat{H}_{\text{inter}} = -\sum_{\mathbf{q}_\parallel,\mathbf{k}_\parallel}\sum_{\lambda\neq\lambda'} \mathbf{d}_{\lambda,\lambda'} \cdot \left(\mathrm{i}\omega_{\lambda,\lambda'}\hat{\mathbf{A}}^{\lambda,\lambda'}_{\mathbf{q}_\parallel}\right) \; a^\dagger_{\lambda,\mathbf{k}_\parallel} a_{\lambda',\mathbf{k}_\parallel+\mathbf{q}_\parallel}. \tag{10.160}$$

For resonant excitation close to the band gap, this relation can be shown to yield approximately identical results as Eq. (10.157), if transformed into the Bloch picture. More explicitly, for a two-band system and using Eq. (10.119), we obtain

$$\hat{H}_{\text{inter}} = -\frac{1}{\sqrt{\mathsf{S}}}\sum_{\mathbf{q},\mathbf{k}_\parallel}\sum_{\lambda} \mathrm{i}\mathsf{E}_{\mathbf{q}}\frac{\omega_{\lambda,\bar{\lambda}}}{\omega_{\mathbf{q}}}\mathbf{d}_{\lambda,\bar{\lambda}} \cdot \left[\mathbf{u}^{\lambda,\bar{\lambda}}_{\mathbf{q}} B_{\mathbf{q}_\parallel,q_z} + \left(\mathbf{u}^{\bar{\lambda},\lambda}_{-\mathbf{q}_\parallel,q_z}\right)^* B^\dagger_{-\mathbf{q}_\parallel,q_z}\right] \; a^\dagger_{\lambda,\mathbf{k}_\parallel} a_{\bar{\lambda},\mathbf{k}_\parallel+\mathbf{q}_\parallel}, \tag{10.161}$$

where $\bar{\lambda}$ denotes the conduction (c) or valence band (v) for $\lambda = v$ or c, respectively. The overlap of the mode functions with the confinement functions has been defined in analogy to Eq. (10.141).

For optical excitations close to the band gap, it is often sufficient to keep only the resonant terms proportional to $B^\dagger a^\dagger_v a_c$ or $B a^\dagger_c a_v$. If we furthermore approximate $\omega_{\mathbf{q}} \approx \omega_{c,v} = -\omega_{v,c}$, we obtain

$$\begin{aligned}
\hat{H}_{\text{inter}} &= -\frac{1}{\sqrt{\mathsf{S}}}\sum_{\mathbf{q},\mathbf{k}_\parallel} \mathrm{i}\mathsf{E}_{\mathbf{q}}\Big[\mathbf{d}_{cv} \cdot \mathbf{u}^{c,v}_{\mathbf{q}} B_{\mathbf{q}_\parallel,q_z} a^\dagger_{c,\mathbf{k}_\parallel} a_{v,\mathbf{k}_\parallel+\mathbf{q}_\parallel} \\
&\quad - \mathbf{d}^*_{cv} \cdot (\mathbf{u}^{c,v}_{-\mathbf{q}_\parallel,q_z})^* B^\dagger_{-\mathbf{q}_\parallel,q_z} \; a^\dagger_{v,\mathbf{k}_\parallel} a_{c,\mathbf{k}_\parallel+\mathbf{q}_\parallel}\Big] \\
&= -\mathrm{i}\hbar\sum_{\mathbf{q},\mathbf{k}_\parallel}\Big[\mathsf{F}_{\mathbf{q}} B_{\mathbf{q}_\parallel,q_z} a^\dagger_{c,\mathbf{k}_\parallel} a_{v,\mathbf{k}_\parallel+\mathbf{q}_\parallel} + \text{h.c.}\Big] \\
&= -\mathrm{i}\hbar\sum_{\mathbf{q}_\parallel,\mathbf{k}_\parallel}\Big[B_{\mathbf{q}_\parallel,\Sigma} \; a^\dagger_{c,\mathbf{k}_\parallel} a_{v,\mathbf{k}_\parallel+\mathbf{q}_\parallel} + \text{h.c.}\Big],
\end{aligned} \tag{10.162}$$

with the matrix element

$$\mathsf{F}_{\mathbf{q}} = \frac{1}{\sqrt{\mathsf{S}}}\frac{1}{\hbar}\,\mathsf{E}_{\mathbf{q}}\mathbf{d}_{cv} \cdot \mathbf{u}^{c,v}_{\mathbf{q}} \tag{10.163}$$

and the collective photon operator

$$B_{\mathbf{q}_\parallel,\Sigma} = \sum_{q_z} \mathsf{F}_{\mathbf{q}_\parallel,q_z} B_{\mathbf{q}_\parallel,q_z}. \tag{10.164}$$

The Hamiltonian, Eq. (10.161), is later used as a starting point for computing optical spectra with dominant interband transitions while the original form in the $(\mathbf{A}\cdot\mathbf{p})$-picture is advantageous for the study of intraband excitations in the THz frequency range. At this point, we would like to remark that a more

rigorous derivation of Eq. (10.162) is possible [7]. It can be shown that in dipole approximation the relation between $(\mathbf{A}\cdot\mathbf{p})$ and $(\mathbf{E}\cdot\mathbf{r})$ picture is given by a unitary transformation. The additional approximations necessary in our derivation are due to the fact that the interpretation of the quantum number $\mathbf{k}$ changes; in our case, $\hbar\mathbf{k}$ labels the canonical momentum of the particle while in the true $(\mathbf{E}\cdot\mathbf{r})$ picture it labels the kinetic momentum. Thus, it is to be expected that also interband transitions look slightly different in both cases.

10.3.4 Phonon–Carrier Interaction

The coupling of electrons to lattice vibrations follows from the third line of the general Hamiltonian, Eq. (10.113). If we use the quantized form of the ion displacement $\Delta\hat{\mathbf{R}}_n$ according to Eq. (10.101), the interaction Hamiltonian becomes

$$\begin{aligned}\hat{H}_{\text{ph-e}} &= -\int \hat{\Psi}^\dagger(\mathbf{r})\left[\sum_n \nabla U\left(\mathbf{r}-\mathbf{R}_n^0\right)\cdot\Delta\hat{\mathbf{R}}_n\right]\hat{\Psi}(\mathbf{r})\mathrm{d}^3r \\ &= -\int \hat{\Psi}^\dagger(\mathbf{r})\left[\sum_{n,\mathbf{p}} \nabla U\left(\mathbf{r}-\mathbf{R}_n^0\right)\cdot\left(D_\mathbf{p}\mathbf{e}_\mathbf{p}-D^\dagger_{-\mathbf{p}}\mathbf{e}_{-\mathbf{p}}\right)\right. \\ &\quad\times\left. \mathrm{i}\sqrt{\frac{\hbar}{2NM\Omega_\mathbf{p}}}\mathrm{e}^{\mathrm{i}\mathbf{R}_n^0\cdot\mathbf{p}}\right]\hat{\Psi}(\mathbf{r})\mathrm{d}^3r \\ &= \sum_{\lambda,\mathbf{k}_\parallel,\lambda',\mathbf{k}'_\parallel} I^{\lambda,\mathbf{k}_\parallel}_{\lambda',\mathbf{k}'_\parallel}\big|_{\text{ph-e}}\, a^\dagger_{\lambda,\mathbf{k}_\parallel} a_{\lambda',\mathbf{k}'_\parallel}. \end{aligned} \tag{10.165}$$

As in Eq. (10.116), the last form is determined by the matrix elements between the Bloch electrons. By expressing the field operators via Eq. (10.39), we obtain

$$\begin{aligned} I^{\lambda,\mathbf{k}_\parallel}_{\lambda',\mathbf{k}'_\parallel}\big|_{\text{ph-e}} &= -\sum_{n,\mathbf{p}}\frac{1}{S}\int \mathrm{e}^{\mathrm{i}(\mathbf{k}'_\parallel-\mathbf{k}_\parallel)\cdot\mathbf{r}_\parallel}\ \mathrm{e}^{\mathrm{i}\mathbf{R}_n^0\cdot\mathbf{p}}\xi^*_\lambda(z)\xi_{\lambda'}(z)\,u^*_{\lambda,\mathbf{k}_\parallel}(\mathbf{r})\,u_{\lambda',\mathbf{k}'_\parallel}(\mathbf{r}) \\ &\quad\times\nabla U\left(\mathbf{r}-\mathbf{R}_n^0\right)\cdot\mathrm{i}\sqrt{\frac{\hbar}{2NM\Omega_\mathbf{p}}}\left(D_\mathbf{p}\mathbf{e}_\mathbf{p}-D^\dagger_{-\mathbf{p}}\mathbf{e}_{-\mathbf{p}}\right)\mathrm{d}^3r. \end{aligned} \tag{10.166}$$

As for the light–mater interaction, we start the evaluation of this matrix element by separating the length scales after we perform the integral over each unit cell. This step leads to

$$
\begin{aligned}
I^{\lambda,\mathbf{k}_\parallel}_{\lambda',\mathbf{k}'_\parallel}|_{\mathrm{ph-e}} &= -\sum_{n,j,\mathbf{p}} \frac{1}{\mathsf{S}} \int_{v_j} \mathrm{e}^{\mathrm{i}(\mathbf{k}'_\parallel-\mathbf{k}_\parallel)\cdot\mathbf{r}_\parallel}\ \mathrm{e}^{\mathrm{i}\mathbf{R}^0_n\cdot\mathbf{p}} \xi^*_\lambda(z)\xi_{\lambda'}(z) \\
&\qquad \times \quad u^*_{\lambda,\mathbf{k}_\parallel}(\mathbf{r})\ \nabla U\left(\mathbf{r}-\mathbf{R}^0_n\right) u_{\lambda',\mathbf{k}'_\parallel}(\mathbf{r})\mathrm{d}^3 r \\
&\qquad \times \mathrm{i}\sqrt{\frac{\hbar}{2NM\Omega_\mathbf{p}}}\left(D_\mathbf{p}\mathbf{e}_\mathbf{p} - D^\dagger_{-\mathbf{p}}\mathbf{e}_{-\mathbf{p}}\right) \\
&= -\sum_{n,j,\mathbf{p}} \frac{1}{\mathsf{S}} \mathrm{e}^{\mathrm{i}(\mathbf{k}'_\parallel-\mathbf{k}_\parallel)\cdot\mathbf{R}^0_{\parallel,j}}\ \mathrm{e}^{\mathrm{i}\mathbf{R}^0_n\cdot\mathbf{p}} \xi^*_\lambda(Z^0_j)\xi_{\lambda'}(Z^0_j) v_j \\
&\qquad \times \quad \frac{1}{v_j}\int_{v_j} u^*_{\lambda,\mathbf{k}_\parallel}(\mathbf{r})\ \nabla U\left(\mathbf{r}-\mathbf{R}^0_n\right)\ u_{\lambda',\mathbf{k}'_\parallel}(\mathbf{r})\mathrm{d}^3 r \\
&\qquad \times \quad \mathrm{i}\sqrt{\frac{\hbar}{2NM\Omega_\mathbf{p}}}\left(D_\mathbf{p}\mathbf{e}_\mathbf{p} - D^\dagger_{-\mathbf{p}}\mathbf{e}_{-\mathbf{p}}\right) \\
&= -\sum_{n,j,\mathbf{p}} \frac{1}{\mathsf{S}} \mathrm{e}^{\mathrm{i}(\mathbf{k}'_\parallel-\mathbf{k}_\parallel)\cdot\mathbf{R}^0_{\parallel,j}} \mathrm{e}^{\mathrm{i}\mathbf{R}^0_n\cdot\mathbf{p}} \xi^*_\lambda(Z^0_j)\xi_{\lambda'}(Z^0_j) v_0 \\
&\qquad \times \quad \frac{1}{v_0}\int_{v_0} u^*_{\lambda,\mathbf{k}_\parallel}(\mathbf{r}')\ \nabla U\left(\mathbf{r}'+\mathbf{R}^0_j-\mathbf{R}^0_n\right)\ u_{\lambda',\mathbf{k}'_\parallel}(\mathbf{r}')\mathrm{d}^3 r' \\
&\qquad \times \quad \mathrm{i}\sqrt{\frac{\hbar}{2NM\Omega_\mathbf{p}}}\left(D_\mathbf{p}\mathbf{e}_\mathbf{p} - D^\dagger_{-\mathbf{p}}\mathbf{e}_{-\mathbf{p}}\right),
\end{aligned}
\tag{10.167}
$$

where the second step is obtained when we take into account that the plane-wave parts and the confinement functions are practically constant over the unit cell. The last step follows after we use a change of integration variable $\mathbf{r} = \mathbf{r}' + \mathbf{R}^0_j$ and note that the Bloch functions are lattice periodic, i.e., $u_{\lambda',\mathbf{k}'_\parallel}(\mathbf{r}' + \mathbf{R}^0_j) = u_{\lambda',\mathbf{k}'_\parallel}(\mathbf{r}')$. Since $\nabla U(\mathbf{r}' + \mathbf{R})$ depends on both microscopic $(\mathbf{r}')$ and mesoscopic $(\mathbf{R})$ scales, it is convenient to introduce a Fourier expansion on the macroscopic scale:

$$
\begin{aligned}
U_\mathbf{q}(\mathbf{r}) &= \sum_n U\left(\mathbf{r}+\mathbf{R}^0_n\right)\mathrm{e}^{-\mathrm{i}\mathbf{q}\cdot\mathbf{R}^0_n}, \\
U\left(\mathbf{r}+\mathbf{R}^0_n\right) &= \frac{1}{N}\sum_\mathbf{q} U_\mathbf{q}(\mathbf{r}) e^{+\mathrm{i}\mathbf{q}\cdot\mathbf{R}^0_n}.
\end{aligned}
\tag{10.168}
$$

As a result, any $U(\mathbf{r} + \mathbf{R}^0_n)$ can be expressed via its microscopic part $U_\mathbf{q}(\mathbf{r})$ times the mesoscopically varying envelope $\mathrm{e}^{\mathrm{i}\mathbf{q}\cdot\mathbf{R}^0_n}$. Using this separation, we find

$$
\begin{aligned}
\nabla_\mathbf{r} U\left(\mathbf{r}+\mathbf{R}^0_n\right) &= \nabla_\mathbf{R} U\left(\mathbf{r}+\mathbf{R}^0_n\right) \\
&= \frac{1}{N}\sum_\mathbf{q} U_\mathbf{q}(\mathbf{r})\ \nabla_\mathbf{R}\mathrm{e}^{\mathrm{i}\mathbf{q}\cdot\mathbf{R}^0_n} = \frac{1}{N}\sum_\mathbf{q} U_\mathbf{q}(\mathbf{r})\ \mathrm{i}\mathbf{q}\mathrm{e}^{\mathrm{i}\mathbf{q}\cdot\mathbf{R}^0_n}.
\end{aligned}
\tag{10.169}
$$

If this result is inserted to Eq. (10.167), we obtain

$$
\begin{aligned}
I^{\lambda,\mathbf{k}_{\|}}_{\lambda',\mathbf{k}'_{\|}}\big|_{\text{ph-e}} = & \sum_{n,j}\sum_{\mathbf{p},\mathbf{q}} \frac{1}{\mathsf{S}} e^{i(\mathbf{k}'_{\|}-\mathbf{k}_{\|}+\mathbf{q}_{\|})\cdot\mathbf{R}^0_{\|,j}} \frac{1}{\mathsf{N}} e^{i\mathbf{R}^0_n\cdot(\mathbf{p}-\mathbf{q})} \xi^*_{\lambda}(Z^0_j)\, e^{iq_z Z^0_j}\, \xi_{\lambda'}(Z^0_j)\, v_j \\
& \times \quad \frac{1}{v_0}\int_{v_0} u^*_{\lambda,\mathbf{k}_{\|}}(\mathbf{r}')\, U_{\mathbf{q}}(\mathbf{r}')\, u_{\lambda',\mathbf{k}'_{\|}}(\mathbf{r}')\mathrm{d}^3r' \\
& \times \quad \sqrt{\frac{\hbar}{2NM\Omega_{\mathbf{p}}}}\Big(D_{\mathbf{p}}\mathbf{q}\cdot\mathbf{e}_{\mathbf{p}} - D^{\dagger}_{-\mathbf{p}}\mathbf{q}\cdot\mathbf{e}_{-\mathbf{p}}\Big) \\
= & \sum_{\mathbf{p},\mathbf{q}}\frac{1}{\mathsf{S}}\int e^{i(\mathbf{k}'_{\|}-\mathbf{k}_{\|}+\mathbf{q}_{\|})\cdot\mathbf{R}_{\|}}\mathrm{d}^2\mathbf{R}_{\|} \\
& \times \quad \frac{1}{N}\sum_n e^{i\mathbf{R}^0_n\cdot(\mathbf{p}-\mathbf{q})}\int \xi^*_{\lambda}(Z)\, e^{iq_z Z}\, \xi_{\lambda'}(Z)\mathrm{d}Z \\
& \times \quad \frac{1}{v_0}\int_{v_0} u^*_{\lambda,\mathbf{k}_{\|}}(\mathbf{r}')\, U_{\mathbf{q}}(\mathbf{r}')\, u_{\lambda',\mathbf{k}'_{\|}}(\mathbf{r}')\mathrm{d}^3r' \\
& \times \quad \sqrt{\frac{\hbar}{2NM\Omega_{\mathbf{p}}}}\Big(D_{\mathbf{p}}\mathbf{q}\cdot\mathbf{e}_{\mathbf{p}} - D^{\dagger}_{-\mathbf{p}}\mathbf{q}\cdot\mathbf{e}_{-\mathbf{p}}\Big),
\end{aligned}
\tag{10.170}
$$

where the j-sums are converted into integrals since the unit-cell volume $v_0 = \mathrm{d}^2\mathbf{R}_{\|}\, \mathrm{d}Z$ can be considered to be infinitesimal on the mesoscopic scale. Now, the integral over $\mathbf{R}_{\|}$ produces $\delta_{\mathbf{k}'_{\|},\mathbf{k}_{\|}-\mathbf{q}_{\|}}$ while the sum over n leads to $\delta_{\mathbf{q},\mathbf{p}}$. We define a deformation potential matrix element

$$
F^{\,\mathbf{k}_{\|},\lambda}_{\,\mathbf{k}'_{\|},\lambda'}(\mathbf{q}) \equiv \frac{1}{v_0}\int_{v_0} u^*_{\lambda,\mathbf{k}_{\|}}(\mathbf{r}')\, U_{\mathbf{q}}(\mathbf{r}')\, u_{\lambda',\mathbf{k}'_{\|}}(\mathbf{r}')\mathrm{d}^3r', \tag{10.171}
$$

where usually only intraband ($\lambda = \lambda'$) contributions are needed for phonon–electron interaction, since the phonon energy is typically orders of magnitude smaller than the interband transition energy, Furthermore, the microscopic integral, Eq. (10.171), is often approximated by its band-index-dependent long-wavelength deformation constant such that

$$
F^{\,\mathbf{k}_{\|},\lambda}_{\,\mathbf{k}'_{\|},\lambda'}(\mathbf{q}) = F^{\lambda}\delta_{\lambda,\lambda'}. \tag{10.172}
$$

For practical purposes, the explicit value of F^{λ} is often determined experimentally.

If we define a confinement function

$$
g^{\lambda}_{q_z} \equiv \int e^{iq_z Z}\, |\xi_{\lambda'}(Z)|^2\mathrm{d}z, \tag{10.173}
$$

we can write

$$I^{\lambda,\mathbf{k}_\parallel}_{\lambda',\mathbf{k}'_\parallel}|_{\mathrm{ph-e}} = \delta_{\lambda,\lambda'} \sum_{\mathbf{p}} \delta_{\mathbf{k}'_\parallel,\mathbf{k}_\parallel-\mathbf{p}_\parallel}\, g^\lambda_{p_z} (\mathbf{p}\cdot\mathbf{e}_\mathbf{p}) F^\lambda \sqrt{\frac{\hbar}{2NM\Omega_\mathbf{p}}} \Big(D_\mathbf{p} + D^\dagger_{-\mathbf{p}} \Big). \tag{10.174}$$

For transversal phonons, $\mathbf{p}\cdot\mathbf{e}_\mathbf{p}$ vanishes since then $\mathbf{p}$ and $\mathbf{e}_\mathbf{p}$ are orthogonal. For longitudinal phonons, $\mathbf{p}$ and $\mathbf{e}_\mathbf{p} = \mathbf{p}/|\mathbf{p}|$ point to the same direction such that $\mathbf{p}\cdot\mathbf{e}_\mathbf{p} = (-\mathbf{p})\cdot\mathbf{e}_{-\mathbf{p}} = |\mathbf{p}|$. As a result, only longitudinal phonons contribute in the lowest order. By inserting this result to Eq. (10.165), we may express the phonon–electron coupling via

$$\begin{aligned} \hat{H}_{\mathrm{ph-e}} &= \sum_{\lambda,\mathbf{k}_\parallel,\mathbf{p}_\parallel,p_z} G^\lambda_\mathbf{p} \Big[D_{\mathbf{p}_\parallel,p_z} + D^\dagger_{-\mathbf{p}_\parallel,p_z} \Big] a^\dagger_{\lambda,\mathbf{k}_\parallel} a_{\lambda,\mathbf{k}_\parallel-\mathbf{p}_\parallel} \\ &= \sum_{\lambda,\mathbf{k}_\parallel,\mathbf{p}_\parallel} \mathsf{G}^\lambda_{\mathbf{p}_\parallel} a^\dagger_{\lambda,\mathbf{k}_\parallel} a_{\lambda,\mathbf{k}_\parallel-\mathbf{p}_\parallel}. \end{aligned} \tag{10.175}$$

Identifying the mass density ρ of the semiconductor material, the volume of the entire semiconductor system L^3, we write the strength of the phonon interaction:

$$G^\lambda_\mathbf{p} = |\mathbf{p}| F^\lambda g^\lambda_{p_z} \sqrt{\frac{\hbar}{2NM\Omega_\mathbf{p}}} = F^\lambda g^\lambda_{p_z} \sqrt{\frac{\hbar|\mathbf{p}|}{2c_{\mathrm{LA}}\rho\mathsf{L}^3}}, \tag{10.176}$$

which is expressed via the velocity of sound c_{LA} for the longitudinal acoustic phonons.

The collective phonon field is then

$$\mathsf{G}^\lambda_{\mathbf{p}_\parallel} \equiv \sum_{p_z} G^\lambda_\mathbf{p} \Big[D_{\mathbf{p}_\parallel,p_z} + D^\dagger_{-\mathbf{p}_\parallel,p_z} \Big], \tag{10.177}$$

which defines the $\mathbf{p}_\parallel$-dependent component in analogy to Eq. (10.119). We observe that the in-plane momentum is conserved in the process where a phonon is either emitted or absorbed, while the electron undergoes an intraband transition. This fact is illustrated in the left part of Fig. 10.2. The consequences of momentum conservation in different dimensional systems follow in the same way as for the light–matter interaction.

10.3.5 Coulomb Interaction

The last term in the general Hamiltonian (10.110) describes the Coulomb interaction among the Bloch electrons. Implementing the basis function (10.39), the Coulomb Hamiltonian can be expressed as

$$
\begin{aligned}
\hat{H}_\mathrm{C} &= \frac{1}{2}\int \hat{\Psi}^\dagger(\mathbf{r})\hat{\Psi}^\dagger(\mathbf{r}')\ V(\mathbf{r}-\mathbf{r}')\ \hat{\Psi}(\mathbf{r}')\hat{\Psi}(\mathbf{r})\mathrm{d}^3r\ \mathrm{d}^3r' \\
&= \frac{1}{2}\frac{1}{\mathsf{S}^2}\sum_{\mathbf{k}_\parallel,\lambda}\sum_{\mathbf{k}'_\parallel,\lambda'}\sum_{\mathbf{p}_\parallel,\nu}\sum_{\mathbf{p}'_\parallel,\nu'} a^\dagger_{\lambda,\mathbf{k}_\parallel} a^\dagger_{\nu,\mathbf{p}_\parallel} a_{\nu',\mathbf{p}'_\parallel} a_{\lambda',\mathbf{k}'_\parallel} \\
&\times \quad \iint \mathrm{e}^{\mathrm{i}\left[(\mathbf{k}'_\parallel-\mathbf{k}_\parallel)\cdot\mathbf{r}_\parallel+(\mathbf{p}'_\parallel-\mathbf{p}_\parallel)\cdot\mathbf{r}'_\parallel\right]} \xi^*_\lambda(z)\xi^*_\nu(z')\ V(\mathbf{r}-\mathbf{r}')\xi_{\nu'}(z')\xi_{\lambda'}(z) \\
&\times \quad u^*_{\lambda,\mathbf{k}_\parallel}(\mathbf{r})\ u^*_{\nu,\mathbf{p}_\parallel}(\mathbf{r}')u_{\nu',\mathbf{p}'_\parallel}(\mathbf{r}')u_{\lambda',\mathbf{k}'_\parallel}(\mathbf{r})\ \mathrm{d}^3r\ \mathrm{d}^3r' \\
&= \frac{1}{2}\sum_{\mathbf{k}_\parallel,\lambda}\sum_{\mathbf{k}'_\parallel,\lambda'}\sum_{\mathbf{p}_\parallel,\nu}\sum_{\mathbf{p}'_\parallel,\nu'} a^\dagger_{\lambda,\mathbf{k}_\parallel} a^\dagger_{\nu,\mathbf{p}_\parallel} a_{\nu',\mathbf{p}'_\parallel} a_{\lambda',\mathbf{k}'_\parallel} \\
&\times \quad \sum_{j,n}\frac{1}{\mathsf{S}^2}\iint \mathrm{e}^{\mathrm{i}\left[(\mathbf{k}'_\parallel-\mathbf{k}_\parallel)\cdot\mathbf{R}_{\parallel,j}+(\mathbf{p}'_\parallel-\mathbf{p}_\parallel)\cdot\mathbf{R}_{\parallel,n}\right]} \\
&\times \quad \xi^*_\lambda(Z_j)\xi^*_\nu(Z_n)\ V(\mathbf{R}_j-\mathbf{R}_n)\xi_{\nu'}(Z_n)\xi_{\lambda'}(Z_j)\ v_j\ v_n \\
&\times \quad \frac{1}{v_j}\int_{v_j} u^*_{\lambda,\mathbf{k}_\parallel}(\mathbf{r})\ u_{\lambda',\mathbf{k}'_\parallel}(\mathbf{r})\mathrm{d}^3r\ \frac{1}{v_n}\int_{v_n} u^*_{\nu,\mathbf{p}_\parallel}(\mathbf{r}')u_{\nu',\mathbf{p}'_\parallel}(\mathbf{r}')\ \mathrm{d}^3r'\,. \qquad (10.178)
\end{aligned}
$$

The final form is obtained if we assume that the Coulomb potential $V(\mathbf{r})$ as well as the envelope and plane-wave parts vary on a mesoscopic scale. Since a unit-cell volume is infinitesimal on the mesoscopic scale, we convert the sums to integrals:

$$
\begin{aligned}
\hat{H}_\mathrm{C} &= \frac{1}{2}\sum_{\mathbf{k}_\parallel,\lambda}\sum_{\mathbf{k}'_\parallel,\lambda'}\sum_{\mathbf{p}_\parallel,\nu}\sum_{\mathbf{p}'_\parallel,\nu'}\frac{1}{\mathsf{S}^2}\int\int \mathrm{e}^{\mathrm{i}\left[(\mathbf{k}'_\parallel-\mathbf{k}_\parallel)\cdot\mathbf{R}_\parallel+(\mathbf{p}'_\parallel-\mathbf{p}_\parallel)\cdot\mathbf{R}'_\parallel\right]} \\
&\times \quad \xi^*_\lambda(Z)\xi^*_\nu(Z')\ V(\mathbf{R}-\mathbf{R}')\xi_{\nu'}(Z')\xi_{\lambda'}(Z)\ \mathrm{d}^3R\ \mathrm{d}^3R' \\
&\times \quad \langle\lambda,\mathbf{k}_\parallel|\lambda',\mathbf{k}'_\parallel\rangle\ \langle\nu,\mathbf{p}_\parallel|\nu',\mathbf{p}'_\parallel\rangle\ a^\dagger_{\lambda,\mathbf{k}_\parallel} a^\dagger_{\nu,\mathbf{p}_\parallel} a_{\nu',\mathbf{p}'_\parallel} a_{\lambda',\mathbf{k}'_\parallel}, \qquad (10.179)
\end{aligned}
$$

where we use the fact that the microscopic integrations represent projections between two different Bloch vectors. To evaluate the mesoscopic integrals, we express the Coulomb potential via its Fourier expansion:

$$
\begin{aligned}
V(\mathbf{r}) &= \frac{e^2}{4\pi\varepsilon\varepsilon_0|\mathbf{r}|} = \sum_{\mathbf{q}_\parallel,q_z}\frac{e^2}{\varepsilon\varepsilon_0\mathsf{L}^3}\frac{1}{|\mathbf{q}|^2}\mathrm{e}^{\mathrm{i}[\mathbf{q}_\parallel\cdot\mathbf{r}_\parallel+q_z z]} \\
&= \sum_{\mathbf{q}_\parallel}\frac{e^2}{\varepsilon\varepsilon_0\mathsf{L}^2}\mathrm{e}^{\mathrm{i}\mathbf{q}_\parallel\cdot\mathbf{r}_\parallel}\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{\mathrm{e}^{\mathrm{i}q_z z}}{\mathbf{q}_\parallel^2+q_z^2}\mathrm{d}q_z, \qquad (10.180)
\end{aligned}
$$

which follows from changing the q_z sum into an integral according to Eq. (10.64). This integral can be solved analytically with the result

$$V(\mathbf{r}) = \sum_{\mathbf{q}_\parallel} \frac{e^2}{\varepsilon\varepsilon_0 \mathsf{L}^2} \mathrm{e}^{\mathrm{i}\mathbf{q}_\parallel \cdot \mathbf{r}_\parallel} \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\mathrm{e}^{\mathrm{i}q_z z}}{(q_z - \mathrm{i}|\mathbf{q}_\parallel|)(q_z + \mathrm{i}|\mathbf{q}_\parallel|)} \mathrm{d}q_z$$
$$= \sum_{\mathbf{q}_\parallel} \frac{e^2}{2\varepsilon\varepsilon_0 \mathsf{L}^2} \frac{1}{|\mathbf{q}_\parallel|} \mathrm{e}^{\mathrm{i}\mathbf{q}_\parallel \cdot \mathbf{r}_\parallel} \mathrm{e}^{-|\mathbf{q}_\parallel||z|}, \tag{10.181}$$

where the last step follows if we use Cauchy's integral theorem where the integration path can be extended in the complex plane such that the closed contour contains one pole at either $q_z = \mathrm{i}|\mathbf{q}_\parallel|$ or $q_z = -\mathrm{i}|\mathbf{q}_\parallel|$.

We use this result in Eq. (10.179) to obtain

$$\begin{aligned}
\hat{H}_\mathrm{C} =& \frac{1}{2} \sum_{\mathbf{k}_\parallel,\lambda} \sum_{\mathbf{k}'_\parallel,\lambda'} \sum_{\mathbf{p}_\parallel,\nu} \sum_{\mathbf{p}'_\parallel,\nu'} \sum_{\mathbf{q}_\parallel} \\
& \frac{1}{\mathsf{S}} \int \mathrm{e}^{\mathrm{i}(\mathbf{k}'_\parallel - \mathbf{k}_\parallel + \mathbf{q}_\parallel)\cdot \mathbf{R}_\parallel} \, \mathrm{d}^2 R_\parallel \;\; \frac{1}{\mathsf{S}} \int \mathrm{e}^{\mathrm{i}(\mathbf{p}'_\parallel - \mathbf{p}_\parallel - \mathbf{q}_\parallel)\cdot \mathbf{R}'_\parallel} \, \mathrm{d}^2 R'_\parallel \\
\times & \int\!\!\int \xi^*_\lambda(Z)\xi^*_\nu(Z') \frac{e^2}{2\varepsilon\varepsilon_0 \mathsf{L}^2} \frac{1}{|\mathbf{q}_\parallel|} \mathrm{e}^{-|\mathbf{q}_\parallel(Z-Z')|} \xi_{\nu'}(Z')\xi_{\lambda'}(Z) \, \mathrm{d}Z \, \mathrm{d}Z' \\
\times & \langle \lambda, \mathbf{k}_\parallel | \lambda', \mathbf{k}'_\parallel \rangle \langle \nu, \mathbf{p}_\parallel | \nu', \mathbf{p}'_\parallel \rangle \, a^\dagger_{\lambda,\mathbf{k}_\parallel} a^\dagger_{\nu,\mathbf{p}_\parallel} a_{\nu',\mathbf{p}'_\parallel} a_{\lambda',\mathbf{k}'_\parallel} \\
=& \frac{1}{2} \sum_{\mathbf{k}_\parallel,\lambda} \sum_{\mathbf{k}'_\parallel,\lambda'} \sum_{\mathbf{p}_\parallel,\nu} \sum_{\mathbf{p}'_\parallel,\nu'} \delta_{\mathbf{k}'_\parallel,\mathbf{k}_\parallel - \mathbf{q}_\parallel} \delta_{\mathbf{p}'_\parallel,\mathbf{p}_\parallel + \mathbf{q}_\parallel} \frac{e^2}{2\varepsilon\varepsilon_0 \mathsf{L}^2} \frac{1}{|\mathbf{q}_\parallel|} \\
\times & \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \xi^*_\lambda(Z)\xi^*_\nu(Z') \, e^{-|\mathbf{q}_\parallel(Z-Z')|} \xi_{\nu'}(Z')\xi_{\lambda'}(Z) \, \mathrm{d}Z \, \mathrm{d}Z' \\
\times & \langle \lambda, \mathbf{k}_\parallel | \lambda', \mathbf{k}_\parallel - \mathbf{q}_\parallel \rangle \langle \nu, \mathbf{p}_\parallel | \nu', \mathbf{p}_\parallel + \mathbf{q}_\parallel \rangle \\
\times & a^\dagger_{\lambda,\mathbf{k}_\parallel} a^\dagger_{\nu,\mathbf{p}_\parallel} a_{\nu',\mathbf{p}_\parallel + \mathbf{q}_\parallel} a_{\lambda',\mathbf{k}_\parallel - \mathbf{q}_\parallel}
\end{aligned} \tag{10.182}$$

since the integrals over $\mathbf{R}_\parallel$ and $\mathbf{R}'_\parallel$ produce delta functions. The expression (10.182) simplifies further when we notice that

$$\langle \lambda, \mathbf{k}_\parallel | \lambda', \mathbf{k}_\parallel \pm \mathbf{q}_\parallel \rangle = \delta_{\lambda,\lambda'} \tag{10.183}$$

for mesoscopic $\mathbf{q}_\parallel$, as was used also for the light–electron and phonon–electron interactions. Before we enter this result to Eq. (10.182), we define the Coulomb matrix element for a quantum well according to

$$V^{\lambda,\lambda'}_{\mathbf{q}_\parallel} \equiv \frac{e^2}{2\varepsilon\varepsilon_0 \mathsf{L}^2} \frac{1}{|\mathbf{q}_\parallel|} \int\!\!\int |\xi_\lambda(z)|^2 |\xi_{\lambda'}(z')|^2 \; \mathrm{e}^{-|\mathbf{q}_\parallel||z-z'|} \, \mathrm{d}z \, \mathrm{d}z'. \tag{10.184}$$

With these observations and definitions, we cast the Coulomb interaction into its final form:

$$\hat{H}_{\mathrm{C}} = \frac{1}{2} \sum_{\mathbf{k}_{\|},\lambda} \sum_{\mathbf{k}'_{\|},\lambda'} \sum_{\mathbf{q}_{\|} \neq 0} V^{\lambda,\lambda'}_{\mathbf{q}_{\|}} \, a^{\dagger}_{\lambda,\mathbf{k}_{\|}} a^{\dagger}_{\lambda',\mathbf{k}'_{\|}} a_{\lambda',\mathbf{k}'_{\|}+\mathbf{q}_{\|}} a_{\lambda,\mathbf{k}_{\|}-\mathbf{q}_{\|}}, \tag{10.185}$$

where the summation indices have been relabeled. This equation shows that the Coulomb interaction transfers the momentum $\mathbf{q}_{\|}$ between two electrons. Thus, $\hat{H}_{\mathrm{C}}$ is a genuine many-body interaction where the in-plane momentum is conserved at the elementary level. We also notice that the $\mathbf{q}_{\|} = 0$ component would lead to a diverging energy contribution in the Coulomb interaction Hamiltonian. This contribution is fully compensated by the Coulomb self-energy of the charged background of ions. The jellium model for the ionic cores where the ions are treated as a uniform background charge density leads to the cancelation of the divergent term [4]. Hence, the diverging contribution $\mathbf{q}_{\|} = 0$ has to be left out from the Coulomb interactions in order to avoid unphysical features.

10.3.6 Complete System Hamiltonian in Different Dimensions

The derivation of the system Hamiltonian is pretty much independent of the dimensionality of the system. The only major changes result from the differences in the momentum conservation and confinement matrix elements.

Starting point for all quantum-well investigations is the total Hamiltonian

$$\hat{H}^{\mathrm{QW}}_{\mathrm{tot}} = \hat{H}_0 + \hat{H}_{\mathrm{em}} + \hat{H}_{\mathrm{ph}} + \hat{H}_{\mathrm{em-e}} + \hat{H}_{\mathrm{ph-e}} + \hat{H}_{\mathrm{C}}, \tag{10.186}$$

where the different contributions are given by Eqs. (10.50), (10.85), (10.107), (10.156), (10.175), and (10.185). For the study of optical interband transition, also the light–matter interaction in the form of Eq. (10.162) is used instead of Eq. (10.156). All these contributions have been derived for a quantum well with confinement in the z-direction such that the electrons are effectively two-dimensional particles with confinement functions $\xi_{\lambda}(z)$. Due to this limitation, the carrier momenta are two dimensional, while the phonon and photon momenta are three dimensional.

Having performed the detailed derivation for quantum wells, it is straightforward to directly construct the total Hamiltonian for other systems with any given dimension. As an example, we use the three-dimensional bulk Hamiltonian. Since there is no confinement in this system, translational invariance is given in all directions such that the total three-dimensional momentum vector must be conserved in all microscopic processes. Once this is taken into account, the equivalent contributions of Eq. (10.186) for a three-dimensional system are given by

$$\hat{H}_0 = \sum_{\lambda,\mathbf{k}} \epsilon^{\lambda}_{\mathbf{k}} a^{\dagger}_{\lambda,\mathbf{k}} a_{\lambda,\mathbf{k}}, \tag{10.187}$$

$$\hat{H}_{\mathrm{em}} = \sum_{\mathbf{q}} \hbar\omega_{\mathbf{q}} \left[B_{\mathbf{q}}^{\dagger} B_{\mathbf{q}} + \frac{1}{2} \right], \tag{10.188}$$

$$\hat{H}_{\mathrm{ph}} = \sum_{\mathbf{p}} \hbar\Omega_{\mathbf{p}} \left(D_{\mathbf{p}}^{\dagger} D_{\mathbf{p}} + \frac{1}{2} \right), \tag{10.189}$$

$$\begin{aligned}\hat{H}_{\mathrm{em-e}} = \sum_{\lambda,\mathbf{k}} \Bigg[&- \sum_{\mathbf{q}} j_{\lambda}(\mathbf{k},\mathbf{q})\; A_{\mathbf{q}}\;\; a_{\lambda,\mathbf{k}-\frac{\mathbf{q}}{2}}^{\dagger} a_{\lambda,\mathbf{k}+\frac{\mathbf{q}}{2}} \\ &+ \sum_{\mathbf{q},\mathbf{q}'} \frac{Q^2}{2m_0} A_{\mathbf{q},-\mathbf{q}'}^{(2)}\;\; a_{\lambda,\mathbf{k}+\mathbf{q}}^{\dagger} a_{\lambda,\mathbf{k}+\mathbf{q}'} \Bigg] \\ &- \sum_{\mathbf{q},\mathbf{k}} \sum_{\lambda \neq \lambda'} Q \mathbf{p}_{\lambda,\lambda'} \cdot \hat{\mathbf{A}}_{\mathbf{q}}\;\; a_{\lambda,\mathbf{k}}^{\dagger} a_{\lambda',\mathbf{k}+\mathbf{q}}, \end{aligned} \tag{10.190}$$

$$\hat{H}_{\mathrm{ph-e}} = \sum_{\lambda,\mathbf{k},\mathbf{p}} \mathsf{G}_{\mathbf{p}}^{\lambda} a_{\lambda,\mathbf{k}}^{\dagger} a_{\lambda,\mathbf{k}-\mathbf{p}}, \tag{10.191}$$

$$\hat{H}_C = \frac{1}{2} \sum_{\lambda,\mathbf{k}} \sum_{\lambda',\mathbf{k}'} \sum_{\mathbf{q}\neq 0} V_{\mathbf{q}}\; a_{\lambda,\mathbf{k}}^{\dagger} a_{\lambda',\mathbf{k}'}^{\dagger} a_{\lambda',\mathbf{k}'+\mathbf{q}} a_{\lambda,\mathbf{k}-\mathbf{q}}. \tag{10.192}$$

The light–matter interaction part contains a current-matrix element with the three-dimensional carrier momentum

$$j_{\lambda}(\mathbf{k},\mathbf{q}) \equiv \frac{Q\hbar\mathbf{k}}{m_{\lambda}} \cdot \mathbf{e}_{\mathbf{q}}, \tag{10.193}$$

where $\mathbf{e}_{\mathbf{q}}$ implicitly includes the polarization direction of the light field. Since Bloch electrons are not confined in a bulk system, the envelope function is a plane wave, such that the strength of the phonon–electron interaction is given by

$$G_{\mathbf{p}}^{\lambda,3\mathrm{D}} = F^{\lambda} \sqrt{\frac{\hbar|\mathbf{p}|}{2c_{\mathrm{LA}}\rho \mathsf{L}^3}}, \tag{10.194}$$

where the deformation potential F^{λ} contains no confinement integrals. In the same way, the Coulomb-matrix element is now

$$V_{\mathbf{q}} = \frac{e^2}{\varepsilon\varepsilon_0 \mathsf{L}^3} \frac{1}{\mathbf{q}^2}. \tag{10.195}$$

Additionally, the different $\mathbf{q}$ components of the linear and the quadratic vector potential are

$$\hat{\mathrm{A}}_{\mathbf{q}} = \frac{\mathsf{E}_{\mathbf{q}}}{\sqrt{\mathsf{L}^3}\,\omega_{\mathbf{q}}} \left(B_{\mathbf{q}} + B_{-\mathbf{q}}^{\dagger} \right), \tag{10.196}$$

$$A^{(2)}_{\mathbf{q},-\mathbf{q}'} = \frac{\mathsf{E}_{\mathbf{q}}, \mathsf{E}_{\mathbf{q}'}}{\mathsf{L}^3 \omega_{\mathbf{q}} \omega_{\mathbf{q}'}} \left(B_{\mathbf{q}} + B^{\dagger}_{-\mathbf{q}} \right) \left(B_{-\mathbf{q}'} + B^{\dagger}_{\mathbf{q}',} \right) \mathbf{e}_{\mathbf{q}} \cdot \mathbf{e}_{-\mathbf{q}'}, \tag{10.197}$$

respectively. In the same way, the three-dimensional phonon field is obtained from

$$\mathsf{G}^{\lambda}_{\mathbf{p}} \equiv G^{\lambda,3\mathrm{D}}_{\mathbf{p}} \left[D_{\mathbf{p}} + D^{\dagger}_{-\mathbf{p}} \right]. \tag{10.198}$$

The Hamiltonian, Eqs. (10.187), (10.188), (10.189), (10.190), (10.191), and (10.192), can be used as a general starting point for the study of light–matter interaction in a bulk system.

From the Hamiltonian of the two-dimensional quantum wells, it is also straightforward to generate the corresponding Hamiltonian for quantum wires or quantum dots. As major modifications, the confinement is now two or even three dimensional such that the necessary modifications have to be made. Other than that, there is no additional complication involved, so that we skip any detailed presentation of equations.

10.4 Quantum Dynamics and Cluster-Expansion Solution

In all quantum-mechanical theories, the Hamilton operator plays a prominent role. It defines the energy eigenvalues of the system and thus allows for expressing the formal solution directly in the usual Schrödinger wave mechanics. Similarly, minimization procedures applied to the expectation value of the Hamiltonian with certain ansatz wave functions may allow one to approximate the true ground state of a more complicated system where a direct solution is impossible.

In the field of many-body physics, a huge number of interacting particles makes a direct solution impossible as well. On the other hand, many interesting physical properties are determined by and can be computed from the single-particle properties alone. For example, the expectation value $\langle a^{\dagger}_{\lambda,\mathbf{k}} a_{\lambda,\mathbf{k}} \rangle$ describes the probability of finding an electron with crystal momentum $\hbar\mathbf{k}$ in band λ. As we will show later the current density entering Maxwell's equations is fully determined by those microscopic intraband distributions. Similarly, the optical polarization is determined by a single-electron transition amplitude between conduction and valence bands. Therefore, instead of solving the full N electron problem (with N being astronomically high in structures of realistic size), one wants to have a formalism in which *reduced* expectation values can be computed without knowledge of the full many-body state in all its detail.

The optimal method is given by the density matrix formalism in the Heisenberg picture. In this formalism, the statistical operator of the interacting many-body/photon/phonon system is time independent and the dynamics of any

expecation value $\langle \hat{O} \rangle$ of a generic operator $\hat{O}$ is fully described by its Heisenberg equation of motion

$$i\hbar \frac{\partial}{\partial t} \langle \hat{O} \rangle = \langle [\hat{O}, \hat{H}_{\text{tot}}]_{-} \rangle \,, \tag{10.199}$$

where $[\ldots]_{-}$ denotes the commutator. The beauty of the approach is that the density matrix does not have to be known at any stage, as long as one knows the expectation values of interest at the beginning of the computation. In general, many coupled equations for expectation values for different quantum numbers have to be solved simultaneously. However, all equations of the form of Eq. (10.199) are standard differential equations for complex functions.

In Section 10.4.2, we use the total Hamiltonian derived previously in order to calculate the fundamental equations of motion for single operators. From these general equations, one can then generate all relevant equations of motion for complicated operator products. Without approximations, the resulting system of equations is in principle exact. Due to the interactions, however, expectation values of a single operator couple to operator products involving higher order correlations. This infinite hierarchy of equations can in practice be truncated based on physical arguments. For example, trions or biexcitons which are examples of correlated complexes of three of four electrons and holes are much less robust than excitons which in turn are less robust than single-particle electron and hole distributions. Thus, a consistent cluster expansion [9, 10, 11, 12, 13, 14, 15, 16] makes a controlled truncation of the hierarchy problem possible. This truncation which will be treated in Section 10.4.3 results in a closed system of equations which can be solved numerically or (for certain special cases) analytically.

10.4.1 Commutator Properties

The Heisenberg equation of motion, Eq. (10.199), defines a procedure of how to obtain the equations of motion for all operator combinations of interest. Before taking the expectation value, we derive all equations of motion on an operator level. For all operators of interest, we must calculate

$$i\hbar \frac{\hat{\partial}}{\partial t} \hat{O} = [\hat{O}, \hat{H}]_{-} \,. \tag{10.200}$$

In order to simplify the derivation of this equation, it is useful to investigate general properties of the commutator:

$$[\hat{A}, \hat{B}]_{-} = \hat{A}\hat{B} - \hat{B}\hat{A}. \tag{10.201}$$

From the definition, it is evident that the commutator is linear in the second argument, i.e.,

$$\left[\hat{A}, \sum_j c_j \hat{B}_j\right]_- = \sum_j c_j [\hat{A}, \hat{B}_j]_- \tag{10.202}$$

for any complex-valued coefficients c_j. Due to the property

$$[\hat{A}, \hat{B}] = -[\hat{B}, \hat{A}]_-, \tag{10.203}$$

it is also linear in the first argument. Every sum of operator products can thus be treated one by one.

A single operator typically still consists of a product of several electronic, photonic, and phononic operators. Therefore, it is practical to have a recursive scheme of how to derive the commutator relation for a general product operator from more elementary operator. This is possible with the relation

$$\begin{aligned} [\hat{A}, \hat{B}\hat{C}]_- &= \hat{A}\hat{B}\hat{C} - \hat{B}\hat{C}\hat{A} \\ &= \hat{A}\hat{B}\hat{C} - \hat{B}\hat{A}\hat{C} + \hat{B}\hat{A}\hat{C} - \hat{B}\hat{C}\hat{A} \\ &= [\hat{A}, \hat{B}]_-\hat{C} + \hat{B}[\hat{A}, \hat{C}]_-, \end{aligned} \tag{10.204}$$

which is valid for any operators $\hat{A}$, $\hat{B}$, and $\hat{C}$. In the case of Fermions, the same relation can be used up to the second-to-last step. Then, one has to reduce the commutation relation to more elementary anti-commutation relations and therefore use

$$\begin{aligned} [\hat{A}, \hat{B}\hat{C}]_- &= \hat{A}\hat{B}\hat{C} - \hat{B}\hat{C}\hat{A} \\ &= \hat{A}\hat{B}\hat{C} + \hat{B}\hat{A}\hat{C} - \hat{B}\hat{A}\hat{C} - \hat{B}\hat{C}\hat{A} \\ &= [\hat{A}, \hat{B}]_+\hat{C} - \hat{B}[\hat{A}, \hat{C}]_+ \end{aligned} \tag{10.205}$$

in order to make use of the known elementary Fermionic anti-commutation relations of the electronic operators.

Using the property of Eq. (10.203) together with Eqs. (10.204) and (10.205), one can easily prove

$$[\hat{A}\hat{B}, \hat{C}]_- = \hat{A}[\hat{B}, \hat{C}]_- + [\hat{A}, \hat{C}]_-\hat{B}, \tag{10.206}$$

$$[\hat{A}\hat{B}, \hat{C}]_- = \hat{A}[\hat{B}, \hat{C}]_+ - [\hat{A}, \hat{C}]_+\hat{B}. \tag{10.207}$$

As a last point, we remark that there is one additional simplification due to the fact that the Hamiltonian operator is always Hermitian. Thus, once the

operator dynamics for a general operator $\hat{O}$ is known, we also obtain the dynamics for $\hat{O}^{\dagger}$ without any further commutations since we may use the relation

$$i\hbar\frac{\hat{\partial}}{\partial t}\hat{O}^{\dagger} = [\hat{O}^{\dagger}, \hat{H}]_{-} = -[\hat{H}, \hat{O}^{\dagger}]_{-} = -[\hat{O}, \hat{H}^{\dagger}]_{-}^{\dagger} = -[\hat{O}, \hat{H}]_{-}^{\dagger}, \tag{10.208}$$

where the last step follows from the Hermiticity of $\hat{H}$. This observation reduces the amount of explicit commutations to half when the general equation of motion structure is solved.

10.4.2 General Operator Dynamics

Any operator can be obtained as combination from the elementary carrier operators $a_{\lambda,\mathbf{k}_{\|}}$ and $a^{\dagger}_{\lambda,\mathbf{k}_{\|}}$, photon operators $B_{q_z,\mathbf{q}_{\|}}$ and $B^{\dagger}_{q_z,\mathbf{q}_{\|}}$, and phonon operators $D_{p_z,\mathbf{p}_{\|}}$ and $D^{\dagger}_{p_z,\mathbf{p}_{\|}}$ where the carriers have a two-dimensional momentum $\mathbf{k}_{\|}$ within the quantum well, while the three-dimensional photon and phonon momenta are divided into in-plane ($\mathbf{q}_{\|}$, $\mathbf{p}_{\|}$) and perpendicular (q_z, p_z) components. Consequently, our first task is to derive the operator dynamics for these elementary operators. From a technical point of view, we only need to derive the dynamics for the annihilation operators since we may use Eq. (10.208) to generate the corresponding creation-operator dynamics.

In the following investigations, we perform the explicit derivations with the quantum wells since these results can directly be generalized for systems with arbitrary dimensionality. For the most part, we want to study interband excitations and thus use the total Hamiltonian

$$\hat{H}_{\text{tot}} = \hat{H}_0 + \hat{H}_{\text{em}} + \hat{H}_{\text{ph}} + \hat{H}_{\text{inter}} + \hat{H}_{\text{ph-e}} + \hat{H}_{\text{C}} \tag{10.209}$$

of Eq. (10.186) with the light–matter interaction in the form of Eq. (10.162), and the other contributions as derived in Eqs. (10.50), (10.85), (10.107), (10.175), and (10.185).

Due to the linearity of the commutator expressed in Eq. (10.202), we can evaluate the commutators of annihilation operators with the different contributions of Eq. (10.209) term by term. We notice immediately that the carrier operators commute with $\hat{H}_{\text{em}} + \hat{H}_{\text{ph}}$, the photon operators commute with $\hat{H}_0 + \hat{H}_{\text{ph}} + \hat{H}_{\text{ph-e}} + \hat{H}_{\text{C}}$, and the phonon operators commute with $\hat{H}_0 + \hat{H}_{\text{em}} + \hat{H}_{\text{em-e}} + \hat{H}_{\text{C}}$. Thus, we only need to evaluate commutators

$$\begin{aligned} C^{(1)} &\equiv \left[B_{\mathbf{q}}, \hat{H}_{\text{em}} + \hat{H}_{\text{inter}}\right]_{-}, \\ C^{(2)} &\equiv \left[D_{\mathbf{p}}, \hat{H}_{\text{ph}} + \hat{H}_{\text{ph-e}}\right]_{-}, \\ C^{(3)} &\equiv \left[a_{\mathbf{k}}, \hat{H}_0 + \hat{H}_{\text{C}} + \hat{H}_{\text{inter}} + \hat{H}_{\text{ph-e}}\right]_{-} \end{aligned} \tag{10.210}$$

to express the operator dynamics for the elementary operators of semiconductors.

By evaluating the Heisenberg equation of motion using the system Hamiltonian, Eq. (10.209), we obtain the photon-operator dynamics,

$$\mathrm{i}\hbar\frac{\partial}{\partial t}B_{\mathbf{q}_{\|},q_{\perp}} = -\hbar\omega_{\mathbf{q}}\,B_{\mathbf{q}_{\|},q_{\perp}} + \mathrm{i}\hbar\sum_{\lambda,\mathbf{k}_{\|}}\left[\mathsf{F}^{\lambda}_{\mathbf{q}_{\|},q_{\perp}}\right]^{*}\,a^{\dagger}_{\lambda,\mathbf{k}_{\|}}a_{\bar{\lambda},\mathbf{k}_{\|}-\mathbf{q}_{\|}}, \tag{10.211}$$

$$\mathrm{i}\hbar\frac{\partial}{\partial t}B^{\dagger}_{\mathbf{q}_{\|},q_{\perp}} = -\hbar\omega_{\mathbf{q}}\,B^{\dagger}_{\mathbf{q}_{\|},q_{\perp}} + \mathrm{i}\hbar\sum_{\lambda,\mathbf{k}_{\|}}\mathsf{F}^{\lambda}_{\mathbf{q}_{\|},q_{\perp}}\,a^{\dagger}_{\bar{\lambda},\mathbf{k}_{\|}}a_{\lambda,\mathbf{k}_{\|}-\mathbf{q}_{\|}}. \tag{10.212}$$

Similarly, we find for the phonon-operator dynamics,

$$\mathrm{i}\hbar\frac{\partial}{\partial t}D_{\mathbf{p}_{\|},p_{\perp}} = \hbar\Omega_{\mathbf{p}}\,D_{\mathbf{p}_{\|},p_{\perp}} + \hbar\sum_{\lambda,\mathbf{k}_{\|}}\mathsf{G}^{\lambda}_{\mathbf{p}_{\|},p_{\perp}}a^{\dagger}_{\lambda,\mathbf{k}_{\|}}a_{\lambda,\mathbf{k}_{\|}+\mathbf{p}_{\|}}\,, \tag{10.213}$$

$$\mathrm{i}\hbar\frac{\partial}{\partial t}D^{\dagger}_{\mathbf{p}_{\|},p_{\perp}} = -\hbar\Omega_{\mathbf{p}}\,D^{\dagger}_{\mathbf{p}_{\|},p_{\perp}} - \hbar\sum_{\lambda,\mathbf{k}_{\|}}\mathsf{G}^{\lambda}_{\mathbf{p}_{\|},p_{\perp}}a^{\dagger}_{\lambda,\mathbf{k}_{\|}}a_{\lambda,\mathbf{k}_{\|}-\mathbf{p}_{\|}}\,. \tag{10.214}$$

Clearly, both the phonon and photon equations contain couplings to carrier operators. However, due to the simple form of Eqs. (10.211) and (10.213), the photon- and phonon-operator dynamics can be directly integrated to give

$$\begin{aligned}B_{\mathbf{q}_{\|},q_{\perp}}(t) = {} & B_{\mathbf{q}_{\|},q_{\perp}}(0)\mathrm{e}^{-\mathrm{i}\omega_{\mathbf{q}}t} + \\ & + \sum_{\lambda,\mathbf{k}_{\|}}\left[\mathsf{F}^{\lambda}_{\mathbf{q}_{\|},q_{\perp}}\right]^{*}\int_{0}^{t}\,\mathrm{d}u\;a^{\dagger}_{\lambda,\mathbf{k}_{\|}}(u)a_{\bar{\lambda},\mathbf{k}_{\|}+\mathbf{q}_{\|}}(u)\;\mathrm{e}^{-\mathrm{i}\omega_{q}(t-u)},\end{aligned} \tag{10.215}$$

$$\begin{aligned}D_{\mathbf{p}_{\|},p_{\perp}}(t) = {} & D_{\mathbf{p}_{\|},p_{\perp}}(0)\mathrm{e}^{-\mathrm{i}\Omega_{p}t} + \\ & + \mathrm{i}\sum_{\lambda,\mathbf{k}_{\|}}\mathsf{G}^{\lambda}_{\mathbf{p}_{\|},p_{\perp}}\int_{0}^{t}\,\mathrm{d}u\;a^{\dagger}_{\lambda,\mathbf{k}_{\|}}(u)a_{\lambda,\mathbf{k}_{\|}+\mathbf{p}_{\|}}(u)\mathrm{e}^{-\mathrm{i}\Omega_{p}(t-u)}.\end{aligned} \tag{10.216}$$

These results show that both a single photon or a single phonon operator are formally equivalent to a combination of two-carrier operators. Such a combination of the form $a^{\dagger}a$ is often called a *single-particle operator* since it describes the transition of a single electron from one quantum state to another. Since electrons cannot be created or annihilated, expectation values of the form $\langle a\rangle$ must vanish and the single-particle operators form the lowest-order particle operator of interest. In contrast, from a purely formal point of view, already single photon or phonon annihilation or creation operators correspond to single-particle operators according to Eqs. (10.215) and (10.216).

For the consistent truncation of the hierarchy problem later on, it is helpful to classify all correlations in terms of N-particle correlations. A general N-particle operator has the form

$$O_N = B_1^\dagger \ldots B_{N_1}^\dagger D_1^\dagger \ldots D_{N_2}^\dagger a_1^\dagger .. a_{N_3}^\dagger a_{N_3} .. a_1 D_{N_4} .. D_1 B_{N_5} .. B_1, \tag{10.217}$$

with all possible combinations of N_j fulfilling $N_1 + N_2 + N_3 + N_4 + N_5 = N$. According to this classification, H_C, H_P, and H_D correspond to two-particle interactions. We also notice that the pure photon- and phonon-operator dynamics, Eqs. (10.211), (10.212), (10.213), and (10.214), involves only single-particle terms.

We find a more complicated dynamical equation for the carrier operators:

$$\begin{aligned} \mathrm{i}\hbar \frac{\partial}{\partial t} a_{\lambda,\mathbf{k}_\parallel} = {} & \epsilon^\lambda_{\mathbf{k}_\parallel} a_{\lambda,\mathbf{k}_\parallel} + \sum_{\lambda',\mathbf{k}'_\parallel,\mathbf{l}_\parallel} V_{\mathbf{l}_\parallel} a^\dagger_{\lambda',\mathbf{k}'_\parallel+\mathbf{l}_\parallel} a_{\lambda',\mathbf{k}'_\parallel} a_{\lambda,\mathbf{k}_\parallel+\mathbf{l}_\parallel} \\ & - \mathrm{i}\hbar \sum_{\mathbf{q}_\parallel} \left[B^\lambda_{\mathbf{q}_\parallel,\Sigma} - \left(B^{\bar{\lambda}}_{-\mathbf{q}_\parallel,\Sigma} \right)^\dagger \right] a_{\bar{\lambda},\mathbf{k}_\parallel-\mathbf{q}_\parallel} \\ & + \hbar \sum_{\mathbf{p}_\parallel} \left[D^\lambda_{\mathbf{p}_\parallel,\Sigma} + \left(D^\lambda_{-\mathbf{p}_\parallel,\Sigma} \right)^\dagger \right] a_{\lambda,\mathbf{k}_\parallel-\mathbf{p}_\parallel}, \end{aligned} \tag{10.218}$$

$$\begin{aligned} \mathrm{i}\hbar \frac{\partial}{\partial t} a^\dagger_{\lambda,\mathbf{k}_\parallel} = {} & -\epsilon^\lambda_{\mathbf{k}_\parallel} a^\dagger_{\lambda,\mathbf{k}_\parallel} - \sum_{\lambda',\mathbf{k}'_\parallel,\mathbf{l}_\parallel} V_{\mathbf{l}_\parallel} a^\dagger_{\lambda,\mathbf{k}_\parallel+\mathbf{l}_\parallel} a^\dagger_{\lambda',\mathbf{k}'_\parallel} a_{\lambda',\mathbf{k}'_\parallel+\mathbf{l}_\parallel} \\ & + \mathrm{i}\hbar \sum_{\mathbf{q}_\parallel} \left[B^{\bar{\lambda}}_{\mathbf{q}_\parallel,\Sigma} - \left(B^\lambda_{-\mathbf{q}_\parallel,\Sigma} \right)^\dagger \right] a^\dagger_{\bar{\lambda},\mathbf{k}_\parallel+\mathbf{q}_\parallel} \\ & - \hbar \sum_{\mathbf{p}_\parallel} \left[D^\lambda_{\mathbf{p}_\parallel,\Sigma} + \left(D^\lambda_{-\mathbf{p}_\parallel,\Sigma} \right)^\dagger \right] a^\dagger_{\lambda,\mathbf{k}_\parallel+\mathbf{p}_\parallel}, \end{aligned} \tag{10.219}$$

where the collective phonon operator is defined as

$$D_{\mathbf{p}_\parallel,\Sigma} = \sum_{p_z} G_{\mathbf{p}_\parallel,p_z} D_{\mathbf{p}_\parallel,p_z} \tag{10.220}$$

in direct analogy to Eq. (10.164). This is the positive-frequency part of Eq. (10.177) under the assumption of infinitely high confinement such that $G_{\mathbf{p}_\parallel,p_z}$ is band independent.

If we now analyze the structure of Eqs. (10.218) and (10.219) in detail, we notice terms that couple the dynamics of single-carrier operators to (i) three-carrier operators due to the Coulomb interaction, (ii) a combination of a photon and a carrier operator due to the light–matter interaction, as well as

(iii) one phonon and one-carrier operator due to the carrier–phonon interaction. Since photon and phonon operators are formally equivalent to two-carrier operators, all these complicated terms effectively lead to the coupling of one-carrier operators to combinations of three-carrier operators. In general, Eqs. (10.211), (10.212), (10.213), (10.214), and (10.218), (10.219) can be applied directly to derive the dynamics of any generic N-particle operator (10.217).

Equations (10.211), (10.212), (10.213), (10.214), and (10.218), (10.219) are the first step in the infinite hierarchy of equations where the N-particle operator quantity is coupled to $N+1$ operators. Since the equations of motion for expectation values are directly obtained from those of the operators, the expectation values inherit the same hierarchy problem,

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\langle N\rangle = T[\langle N\rangle] + V[\langle N+1\rangle]. \tag{10.221}$$

Here, the functional T results mainly from the non-interacting part of the Hamiltonian while V originates from the interactions. These interactions couple the N-particle expectation value $\langle N\rangle$ to $\langle N+1\rangle$ quantities. Consequently, Eq. (10.221) cannot be closed and we must resort to a systematic truncation scheme in order to obtain controlled approximations.

10.4.3 Cluster Expansion

One successful approach to deal with the hierarchy problem is to use the so-called cluster-expansion scheme [10, 12, 16, 17]. This approach is well established, e.g., in quantum chemistry where it is used to treat the many-body problems related to molecular eigenstates [9, 11, 15]. In semiconductor systems, this method has been used to analyze a variety of many-body and quantum-optical problems [7, 10, 12, 13, 17, 18, 19, 20]. In the following, we first review the basic idea behind the cluster expansion and then discuss specific aspects that are relevant for the investigations of our semiconductor system.

The cluster-expansion method is based on a clear physical principle where one determines all consistent factorizations of an N-particle quantity $\langle N\rangle$ in terms of (i) independent single particles (singlets), (ii) correlated pairs (doublets), (iii) correlated three-particle clusters (triplets), up to (iv) correlated N-particle clusters. If we formally know all expectation values from $\langle 1\rangle$ to $\langle N\rangle$, a specific correlated cluster can be constructed recursively using

$$\begin{aligned}
\langle 2\rangle &= \langle 2\rangle_{\mathrm{S}} + \Delta\langle 2\rangle,\\
\langle 3\rangle &= \langle 3\rangle_{\mathrm{S}} + \langle 1\rangle\Delta\langle 2\rangle + \Delta\langle 3\rangle,\\
\langle N\rangle &= \langle N\rangle_{\mathrm{S}} + \langle N-2\rangle_{\mathrm{S}}\Delta\langle 2\rangle + \langle N-4\rangle_{\mathrm{S}}\Delta\langle 2\rangle\Delta\langle 2\rangle + \cdots\\
&\quad + \langle N-3\rangle_{\mathrm{S}}\Delta\langle 3\rangle + \langle N-5\rangle_{\mathrm{S}}\Delta\langle 2\rangle\Delta\langle 3\rangle + \cdots + \Delta\langle N\rangle.
\end{aligned} \tag{10.222}$$

Here, the quantities with the subscript S denote the singlet contributions and the terms $\Delta\langle J\rangle$ contain the purely correlated parts of the J-particle cluster. In Eq. (10.222), each term includes a sum over all *unique* possibilities to reorganize the N coordinates among singlets, doublets, and so on. The different reorganizations are defined by permutations of operator indices. To guarantee the fundamental indistinguishability of particles, one must add up all the permutations. For Fermionic operators, one has to take a positive sign for even permutations and a negative sign for odd permutations. For Bosons, all permutations are added with a positive sign. This way, all cluster groups in Eq. (10.222) are fully anti-symmetric for the Fermionic carriers and fully symmetric for the Bosonic photon and phonon operators, respectively.

To obtain more insights into the different contributions appearing in Eq. (10.222), we consider first the singlet factorization. For pure carrier-operator terms, we find the *Hartree–Fock factorization*:

$$\left\langle a_1^\dagger \ldots a_N^\dagger a_N \ldots a_1 \right\rangle_{\mathrm{S}} = \sum_\sigma (-1)^\sigma \prod_{j=1}^{N} \left\langle a_j^\dagger a_{\sigma[j]} \right\rangle \tag{10.223}$$

where σ is an element of the permutation group with indices $1, .., N$. Specifically, $\sigma[j]$ defines the mapping of the index j under the permutation σ. In the sum over all permutations, the even permutations lead to $(-1)^\sigma = +1$ while the odd permutations lead to $(-1)^\sigma = -1$. Equation (10.223) can be written in a more compact form by noting that it actually involves the determinant of a matrix, i.e.,

$$M_{j,k} \equiv \left\langle a_j^\dagger a_k \right\rangle, \qquad \left\langle a_1^\dagger \ldots a_N^\dagger a_N \ldots a_1 \right\rangle_{\mathrm{S}} = \det(\mathbf{M}). \tag{10.224}$$

Pure photon or phonon terms or a mixture of them also allow for a simple singlet factorization:

$$\left\langle b_1^\dagger .. b_M^\dagger b_{M+1} .. b_N \right\rangle_{\mathrm{S}} = \left\langle b_1^\dagger \right\rangle .. \left\langle b_M^\dagger \right\rangle \langle b_{M+1} \rangle .. \langle b_N \rangle, \tag{10.225}$$

where $M \leq N$ and b stands for a generic Boson operator (either photon or phonon) identified by its index. Equation (10.225) clearly describes a classical factorization since each expectation value of a single Bosonic operator represents a complex-valued quantity, $\beta_j = \langle b_j \rangle$, such that one simply obtains the product of the different β_j. With this observation, we realize that also the combination of carrier and Boson operators produces a simple singlet contribution which is obtained by replacing each Boson operator by the corresponding classical β_j while the remaining pure carrier part can be factorized using Eq. (10.224).

The systematic cluster expansion is obtained by decomposing any given N-particle quantity into C-particle correlations,

$$\langle N\rangle_{1..C} \equiv \langle N\rangle_S + \langle N\rangle_D + \cdots + \langle N\rangle_C = \sum_{J=1}^{C} \langle N\rangle_J, \tag{10.226}$$

following directly from Eq. (10.222). Here, $\langle N\rangle_S$ contains only singlets, $\langle N\rangle_D$ contains all combinations of doublets but no higher order correlations and so on. The nature of the respective physical problem determines the lowest possible level at which one might truncate the cluster expansion. For example, one clearly needs calculations at least up to the doublet level for electron–hole systems containing bound pairs, i.e., excitons [17, 21, 22]. Increasingly more clusters have to be included if one wants to describe excitonic molecules or even higher correlations [23, 24]. For the light field, the singlet contributions describe the classical part of the field while the quantum fluctuations are determined by the higher order correlations. In many quantum-optical phenomena, decisive contributions result from the doublet correlation terms such as photon number and two-photon absorption correlations.

For many experimentally relevant situations, one can limit the description of the semiconductor system to a phase space where plasma and excitons coexist but higher order cluster are less important. In this regime, the singlet–doublet approximation describes a multitude of microscopic effects via the factorized N-particle expectation value $\langle N\rangle_{SD}$. At this level of approximation, we need to solve the dynamics of all possible singlets $\langle 1\rangle$ and doublets $\Delta\langle 2\rangle$ because then any arbitrary $\langle N\rangle$ consists only of known combinations of single-particle expectation values and two-particle correlations.

To see the general structure of the relevant singlet–doublet equations, we start from Eq. (10.221) and apply the truncation (10.226) up to three-particle correlations (triplets) – i.e., one level higher than a pure singlet–doublet theory since we want to extend some investigations beyond the doublet level. We find the general equation structure:

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\langle 1\rangle = T_1[\langle 1\rangle] + V_{1a}\left[\langle 2\rangle_S\right] + V_{1b}[\Delta\langle 2\rangle], \tag{10.227}$$

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\Delta\langle 2\rangle = T_2[\Delta\langle 2\rangle] + V_{2a}\left[\langle 3\rangle_{SD}\right] + V_{2b}[\Delta\langle 3\rangle], \tag{10.228}$$

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\Delta\langle 3\rangle = T_3[\Delta\langle 3\rangle] + V_3\left[\langle 4\rangle_{SDT}\right], \tag{10.229}$$

where $T_{1(2,3)}$ and $V_{1(2,3)}$ are known functionals defined by the respective Heisenberg equations of motion. In this form, the structure of the singlet and doublet equations is exact while only the triplet dynamics is approximated. Consequently, the hierarchy is systematically truncated resulting in a finite number of coupled equations.

The evaluation of the full singlet–doublet–triplet approximation, Eqs. (10.227), (10.228), and (10.229), is still beyond current numerical capabilities if one wants to study QW or QWI systems. However, one can find clear physical principles to simplify the triplet dynamics (10.229) since it contains two distinct classes of contributions: (i) the microscopic processes describing scattering effects between two-particle correlations and single-particle quantities and (ii) interactions which are responsible for the formation of genuine three-particle correlations like trions. To the first category belong effects where correlated electron–hole pairs scatter with an electron, hole, or phonon. These interactions lead to screening of the Coulomb interaction, dephasing of the coherences, and formation or equilibration of exciton populations [17, 21, 22].

Since the formation of bound three-particle complexes is slow in QWs and QWIs after optical excitations and requires high densities beyond the exciton Mott transition to become relevant [23, 24], we omit genuine three-particle correlations from the analysis. Thus, we end up with a consistent singlet–doublet approach where we treat triplet correlations at the scattering level. This leads us to the general equation structure:

$$
\begin{aligned}
\mathrm{i}\hbar\frac{\partial}{\partial t}\langle 1\rangle &= T_1[\langle 1\rangle] + V_1\left[\langle 2\rangle_{\mathrm{S}}\right] + V_1[\Delta\langle 2\rangle],\\
\mathrm{i}\hbar\frac{\partial}{\partial t}\Delta\langle 2\rangle &= T_2[\Delta\langle 2\rangle] + V_2\left[\langle 3\rangle_{\mathrm{SD}}\right] + G[\langle 1\rangle, \Delta\langle 2\rangle].
\end{aligned}
\tag{10.230}
$$

Here, the functional $G[\langle 1\rangle, \Delta\langle 2\rangle]$ indicates that three-particle correlations are included at the scattering level. The schematical structure of the approximation behind Eq. (10.230) is depicted in Fig. 10.3.

The pure one- and two-particle dynamics can now be obtained from Eq. (10.230) by evaluating the factorizations $\langle 2\rangle_{\mathrm{SD}}$ and $\langle 3\rangle_{\mathrm{SD}}$. For the detailed calculations, we explicitly need the different singlet–doublet factorizations

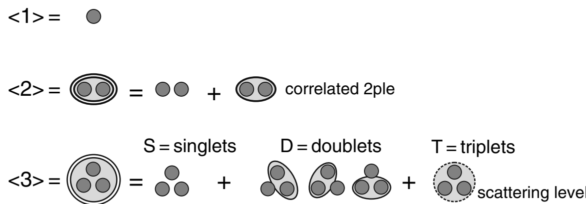


Fig. 10.3 Visualization of the cluster-expansion approach up to the level where singlets and doublets are fully included and the triplets are treated at the scattering level. The first line defines the singlets, the second line shows how the doublets are decomposed into their singlet contributions and the correlated part, and the third line depicts how the triplets are expanded into products of three singlets, products of correlated doublets and singlets, and correlated triplets which are replaced by the scattering-level approximation

which can be obtained directly by applying Eqs. (10.222) and (10.226) for any combination of carrier, photon, or phonon operators. The corresponding derivation of the singlet–doublet dynamics follows a straightforward procedure after one evaluates the explicit forms of the needed factorizations.

10.4.4 Singlet–Doublet Correlations

In practice, the equations of motion for all relevant operator products are obtained by applying the operator, Eqs. (10.211), (10.211), (10.213), (10.214) and (10.218), (10.219), taking the expectation value of the resulting equations of motion, and truncating the higher order N-particle terms according to the singlet–doublet approximation. Here, we collect and present the relevant singlet and doublet correlations which occur during the factorization. Since the single-t–doublet level forms a closed set of equations, any expectation value for the coupled electron–photon–phonon system in singlet–doublet approximation can be expressed in terms of a finite number of different correlations which are briefly presented here. In the later sections where we study the physical examples, we present the resulting equations of motion in explicit notation. The full details of the derivations and the use of an abstract and compactifying implicit notation is described in Kira and Koch [1].

The possible singlet terms are given by

$$\langle 1 \rangle = \left\{ \left\langle a^{\dagger}_{\lambda,\mathbf{k}_{\|}} a_{\lambda',\mathbf{k}'_{\|}} \right\},\ \langle B_{\mathbf{q}_{\|},q_{\perp}} \rangle,\ \text{or}\ \langle D_{\mathbf{p}_{\|},p_{\perp}} \rangle \right\} \tag{10.231}$$

describing electronic transition amplitudes as well as coherent photon or phonon expectation values. Generally, the momenta $\mathbf{k}_{\|}$ and $\mathbf{k}'_{\|}$ can be different. However, in this manuscript, we only consider situations where the system is excited with a homogeneous external light pulse propagating perpendicular to the QW structure. In this configuration, all quantities are homogeneous and the corresponding two-point expectation values vanish for $\mathbf{k}_{\|} \neq \mathbf{k}'_{\|}$, $\mathbf{q}_{\|} \neq \mathbf{0}$, and $\mathbf{p}_{\|} \neq \mathbf{0}$. Thus, we can only have a difference in the band or spin index in the terms $\langle a^{\dagger}_{\lambda,\mathbf{k}_{\|}} a_{\lambda',\mathbf{k}'_{\|}} \rangle$, i.e.,

$$\left\langle a^{\dagger}_{\lambda,\mathbf{k}_{\|}} a_{\lambda',\mathbf{k}'_{\|}} \right\rangle = \delta_{\mathbf{k}_{\|},\mathbf{k}'_{\|}} \left\langle a^{\dagger}_{\lambda,\mathbf{k}_{\|}} a_{\lambda',\mathbf{k}_{\|}} \right\rangle. \tag{10.232}$$

In the same way, the homogeneous coherent light and phonon fields have

$$\langle B_{\mathbf{q}_{\|},q_{\perp}} \rangle = \delta_{\mathbf{q}_{\|},\mathbf{0}} \langle B_{\mathbf{0},q_{\perp}} \rangle, \quad \langle D_{\mathbf{p}_{\|},p_{\perp}} \rangle = \delta_{\mathbf{p}_{\|},\mathbf{0}} \langle D_{\mathbf{0},p_{\perp}} \rangle. \tag{10.233}$$

As we have seen in Section 10.4.2 and in particular in the carrier equations, Eqs. (10.218) and (10.219), the single-particle terms are coupled to doublets. As a consequence of the translational invariance under homogeneous excitation,

the combined in-plane momentum of all creation operators has to be equal to that of the annihilation operators in all correlation terms. For two-particle carrier correlations, we therefore have to demand that

$$\Delta\left\langle a^{\dagger}_{\lambda,\mathbf{k}_{\|}}\, a^{\dagger}_{\nu,\mathbf{k}'_{\|}}\, a_{\nu',\mathbf{k}''_{\|}}\, a_{\lambda',\mathbf{k}'''_{\|}}\right\rangle = \delta_{\mathbf{k}_{\|}+\mathbf{k}'_{\|},\mathbf{k}''_{\|}+\mathbf{k}'''_{\|}}\Delta\left\langle a^{\dagger}_{\lambda,\mathbf{k}_{\|}}\, a^{\dagger}_{\nu,\mathbf{k}'_{\|}}\, a_{\nu',\mathbf{k}''_{\|}}\, a_{\lambda',\mathbf{k}'''_{\|}}\right\rangle$$

$$\equiv \delta_{\mathbf{k}''_{\|},\mathbf{k}'_{\|}+\mathbf{q}_{\|}}\delta_{\mathbf{k}'''_{\|},\mathbf{k}_{\|}-\mathbf{q}_{\|}}\Delta\left\langle a^{\dagger}_{\lambda,\mathbf{k}_{\|}}\, a^{\dagger}_{\nu,\mathbf{k}'_{\|}}\, a_{\nu',\mathbf{k}'_{\|}+\mathbf{q}_{\|}}\, a_{\lambda',\mathbf{k}_{\|}-\mathbf{q}_{\|}}\right\rangle, \qquad (10.234)$$

where the last step defines the momentum condition with the help of a new momentum $\mathbf{q}_{\|}$. This is performed because this identification simply presents the general form of the two-particle carrier correlations as they appear in the further derivations. In our subsequent calculations, we will often identify $\mathbf{q}_{\|}$ as the center-of-mass momentum of the correlated two-particle entities.

For later use, we introduce an abbreviation

$$c^{\mathbf{q}_{\|},\mathbf{k}'_{\|},\mathbf{k}_{\|}}_{\lambda,\nu;\nu',\lambda'} \equiv \Delta\left\langle a^{\dagger}_{\lambda,\mathbf{k}_{\|}}\, a^{\dagger}_{\nu,\mathbf{k}'_{\|}}\, a_{\nu',\mathbf{k}'_{\|}+\mathbf{q}_{\|}}\, a_{\lambda',\mathbf{k}_{\|}-\mathbf{q}_{\|}}\right\rangle \qquad (10.235)$$

for the generic two-particle correlations. Here, the superscripts are arranged such that $\mathbf{q}_{\|}$ indicates the momentum transfer, whereas $\mathbf{k}'_{\|}$ and $\mathbf{k}_{\|}$ are the momentum indices of the second and first creation operator, respectively. The combination of creation and destruction operators in Eq. (10.235) shows that total momentum conservation is satisfied for all band indices $(\lambda, \nu, \nu', \lambda')$.

The doublet correlations with mixed combinations of carrier-, photon-, and phonon operators obey the same conservation law for the in-plane momentum as the pure carrier correlations. Thus, we find that only the combinations

$$\Delta\langle 2\rangle_{\text{mix}} = \Big\{\Delta\left\langle B^{\dagger}_{\mathbf{q}_{\|},q_{\perp}}\, a^{\dagger}_{\nu,\mathbf{k}_{\|}-\mathbf{q}_{\|}}\, a_{\nu',\mathbf{k}_{\|}}\right\rangle,\ \ \Delta\left\langle D^{\dagger}_{\mathbf{p}_{\|},p_{\perp}}\, a^{\dagger}_{\nu,\mathbf{k}_{\|}-\mathbf{p}_{\|}}\, a_{\nu',\mathbf{k}_{\|}}\right\rangle,$$

$$\Delta\left\langle B^{\dagger}_{\mathbf{q}_{\|},q_{\perp}}\, D_{\mathbf{q}_{\|},p_{\perp}}\right\rangle,\ \ \Delta\left\langle B_{\mathbf{q}_{\|},q_{\perp}}\, D_{-\mathbf{q}_{\|},p_{\perp}}\right\rangle\Big\} \qquad (10.236)$$

are allowed. Furthermore, the pure photon and phonon correlations assume the generic form

$$\Delta\langle 2\rangle_{\text{bos}} = \Big\{\Delta\left\langle B^{\dagger}_{\mathbf{q}_{\|},q_{\perp}}\, B_{\mathbf{q}_{\|},q'_{\perp}}\right\rangle,\ \ \Delta\langle B_{\mathbf{q}_{\|},q_{\perp}}\, B_{-\mathbf{q}_{\|},q'_{\perp}}\rangle,$$

$$\Delta\left\langle D^{\dagger}_{\mathbf{p}_{\|},p_{\perp}}\, D_{\mathbf{p}_{\|},p'_{\perp}}\right\rangle,\ \ \Delta\langle D_{\mathbf{p}_{\|},p_{\perp}}\, D_{-\mathbf{p}_{\|},p'_{\perp}}\rangle\Big\}. \qquad (10.237)$$

Equations (10.235), (10.236), and (10.237) define the generic two-particle correlations of homogeneous systems.

10.5 Semiconductor Absorption Spectra

Current experiments utilize a large variety of laser sources ranging from ultra-fast sub-picosecond pulsed systems [25, 26, 27, 28] all the way to continuous-wave (cw) operation. One can apply different measurement schemes to detect the excitation-induced changes in the optical response such as light transmission, reflection, and absorption [29, 30, 31, 32, 33, 34, 35] as well as light scattering and wave-mixing signatures [36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46]. This coherent laser excitation approach is widely used not only to explore quantum-mechanical properties of many-body systems but also with the goal to develop practical devices.

Thus, as a first example, we treat the absorption of classical light by semiconductor structures within the formalism developed in the previous sections. We start with the explicit version of the coupled semiconductor-Bloch–Maxwell equations. Those equations can be used to describe the semiconductor response from the linear up to the highly non-linear excitation regime. The exciting pulse is directly modeled in the time regime. On the other hand, for relatively weak pulses, it is often sufficient to restrict oneself to the linear regime. For example in the case of typical pump-probe-type experiments, the system is excited with a strong pump pulse. After the coherent polarization has decayed, electrons and holes still remain excited in the semiconductor system for times up to several nanoseconds. The density decay and the state of the system is then typically probed by a weak probe pulse arriving somewhen after the pump and before the system has returned to its ground state. When one wants to model such a situation, it is sufficient to compute the response in linear order of the probe pulse. Nevertheless, the carriers in the system provide strong density-dependent Coulomb scattering which leads to drastic density-dependent changes in the semiconductor absorption for different pump strengths.

10.5.1 Semiconductor Bloch Equations

An ideal, coherent laser generates a quantum field that is as close as possible to classical light. Therefore, the interaction of laser light and matter can often be described at the semiclassical level. Here, one uses the classical electrodynamic theory in conjunction with a quantum-mechanical approach to analyze the creation, annihilation, and interaction of the different material excitations. In this spirit, we now specialize the general singlet–doublet formalism of Section 10.4 to the case of a classical light field.

Since a coherent state is eigenstate of the photon annihilation operator, the restriction to a classical light field is equivalent to factorizing all photon operators into their singlet form. For example, the light intensity is related to $\langle B^{\dagger} B \rangle = \langle B^{\dagger} \rangle \langle B \rangle$ and mixed operators are factorized according to, e.g., $\langle B^{\dagger} a^{\dagger} a \rangle = \langle B^{\dagger} \rangle \langle a^{\dagger} a \rangle$. In this classical factorization, $B_{\mathbf{q}}$ is taken out of the expectation values as a complex-valued field. Since this field can be decomposed

into an amplitude and a phase, the classical part of the light field is referred to as *coherent* while the remaining quantum-optical fluctuations are incoherent if they exist without the classical field.

Based on the formal equivalence of light and particle correlations, the classical excitations should – in first order – generate single-particle carrier quantities. Thus, we have to solve the full singlet dynamics in order to determine the principal effects of classical optical excitations. In particular, we have to evaluate the dynamics of the microscopic polarization and the carrier occupations during and after the excitation. For this purpose, we introduce a simplifying notation for the microscopic polarization

$$P_{\mathbf{k}_\parallel} \equiv \left\langle a^\dagger_{v,\mathbf{k}_\parallel} a_{c,\mathbf{k}_\parallel} \right\rangle, \tag{10.238}$$

and for the electron and hole occupations

$$f^{\mathrm{e}}_{\mathbf{k}_\parallel} \equiv n^{c}_{\mathbf{k}_\parallel} = \left\langle a^\dagger_{c,\mathbf{k}_\parallel} a_{c,\mathbf{k}_\parallel} \right\rangle, \tag{10.239}$$

$$f^{\mathrm{h}}_{\mathbf{k}_\parallel} \equiv 1 - n^{v}_{\mathbf{k}_\parallel} = 1 - \left\langle a^\dagger_{v,\mathbf{k}_\parallel} a_{v,\mathbf{k}_\parallel} \right\rangle = \left\langle a_{v,\mathbf{k}_\parallel} a^\dagger_{v,\mathbf{k}_\parallel} \right\rangle. \tag{10.240}$$

Using these notations and the elementary equations (10.218) and (10.219), we can derive the general *semiconductor Bloch equations* (SBE) [4, 47]:

$$\mathrm{i}\hbar \frac{\partial}{\partial t} P_{\mathbf{k}_\parallel} = \tilde{\epsilon}_{\mathbf{k}_\parallel} P_{\mathbf{k}_\parallel} - \left[1 - f^{\mathrm{e}}_{\mathbf{k}_\parallel} - f^{\mathrm{h}}_{\mathbf{k}_\parallel}\right] \Omega_{\mathbf{k}_\parallel} + \Gamma^{v,c}_{\mathbf{k}_\parallel} + \Gamma^{\mathrm{QED}}_{v,c;\mathbf{k}_\parallel}, \tag{10.241}$$

$$\hbar \frac{\partial}{\partial t} f^{\mathrm{e}}_{\mathbf{k}_\parallel} = 2\mathrm{Im}\left[P_{\mathbf{k}_\parallel} \Omega^*_{\mathbf{k}_\parallel} + \Gamma^{c,c}_{\mathbf{k}_\parallel} + \Gamma^{\mathrm{QED}}_{c,c;\mathbf{k}_\parallel}\right], \tag{10.242}$$

$$\hbar \frac{\partial}{\partial t} f^{\mathrm{h}}_{\mathbf{k}_\parallel} = 2\mathrm{Im}\left[P_{\mathbf{k}} \Omega^*_{\mathbf{k}_\parallel} - \Gamma^{v,v}_{\mathbf{k}_\parallel} - \Gamma^{\mathrm{QED}}_{v,v;\mathbf{k}_\parallel}\right], \tag{10.243}$$

where we have applied the cluster expansion and separated single- and two-particle terms. In the singlet terms, we introduced the renormalized kinetic electron–hole pair energy and the renormalized Rabi frequency,

$$\begin{aligned} \tilde{\epsilon}_{\mathbf{k}_\parallel} &\equiv \epsilon^{c}_{\mathbf{k}_\parallel} - \epsilon^{v}_{\mathbf{k}_\parallel} - \sum_{\mathbf{k}'_\parallel} V_{\mathbf{k}_\parallel - \mathbf{k}'_\parallel} \left(f^{\mathrm{e}}_{\mathbf{k}'_\parallel} + f^{\mathrm{h}}_{\mathbf{k}'_\parallel}\right), \\ \Omega_{\mathbf{k}_\parallel} &\equiv d_{c,v} \langle E(0,t) \rangle + \sum_{\mathbf{k}'_\parallel} V_{\mathbf{k}_\parallel - \mathbf{k}'_\parallel} P_{\mathbf{k}'_\parallel} \end{aligned} \tag{10.244}$$

respectively.

In Eqs. (10.241), (10.242), and (10.243) the doublet contributions show up as quantum-optical correlations Γ^{QED} and as microscopic scattering terms

$$\Gamma^{\lambda,\lambda'}_{\mathbf{k}_\parallel} \equiv \sum_{\nu,\mathbf{k}'_\parallel,\mathbf{q}_\parallel \neq \mathbf{0}} V_{\mathbf{q}_\parallel} \left[c^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel}_{\lambda,\nu;\nu,\lambda'} - \left(c^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel}_{\lambda',\nu;\nu,\lambda} \right)^* \right]$$
$$+ \sum_{\mathbf{q}_\parallel} \left[\Delta\left\langle Q^{\lambda'}_{\mathbf{q}_\parallel} a^\dagger_{\lambda,\mathbf{k}_\parallel} a_{\lambda',\mathbf{k}_\parallel - \mathbf{q}_\parallel} \right\rangle - \Delta\left\langle \left(Q^{\lambda}_{\mathbf{q}_\parallel} \right)^\dagger a^\dagger_{\lambda,\mathbf{k}_\parallel - \mathbf{q}_\parallel} a_{\lambda',\mathbf{k}_\parallel} \right\rangle \right] \quad (10.245)$$

due to the Coulomb (first line) and the phonon–carrier (second line) interactions. In general, the doublet terms $\Gamma^{\lambda,\lambda'}$ introduce microscopic couplings to the two-particle Coulomb and phonon correlations, which describe dephasing, energy renormalizations, and screening, as well as relaxation of the carrier densities toward steady-state distributions.

The quantum-optical two-particle correlations have the explicit form

$$\Gamma^{\mathrm{QED}}_{\lambda,\lambda';\mathbf{k}_\parallel} \equiv - \sum_{\mathbf{q}_\parallel} \left[\Delta\left\langle E^{\lambda'}_{\mathbf{q}_\parallel} a^\dagger_{\lambda,\mathbf{k}_\parallel} a_{\bar{\lambda}',\mathbf{k}_\parallel - \mathbf{q}_\parallel} \right\rangle - \Delta\left\langle \left(E^{\lambda}_{\mathbf{q}_\parallel} \right)^\dagger a^\dagger_{\bar{\lambda},\mathbf{k}_\parallel - \mathbf{q}_\parallel} a_{\lambda',\mathbf{k}_\parallel} \right\rangle \right]. \quad (10.246)$$

They introduce all combinations of photon-assisted correlations which describe the spontaneous recombination of carriers, entanglement effects, or the generation of densities via quantum-light absorption. Since these effects often play a minor role in typical classical optical experiments, we omit the Γ^{QED} contributions for this discussion.

For a classical light field, the microscopic polarization equation is coupled to the wave equation:

$$\left[\frac{\partial^2}{\partial r_\perp^2} - \frac{n^2(r_\perp)}{c^2} \frac{\partial^2}{\partial t^2} \right] \langle E(r_\perp, t) \rangle = \mu_0 |\xi(r_\perp)|^2 \frac{\partial^2}{\partial t^2} P, \quad (10.247)$$

where the classical light field is directly influenced by the macroscopic optical polarization given by

$$P = \frac{d_{vc}}{\mathsf{S}} \sum_{\mathbf{k}_\parallel} P_{\mathbf{k}_\parallel} + \mathrm{c.c.}, \quad (10.248)$$

with the quantization area S. Since the macroscopic polarization follows from the singlet term $P_{\mathbf{k}_\parallel}$, the wave equation involves only single-particle contributions without the hierarchy problem. We note that an operator version of the wave equation can be derived for a quantized light field as well, starting from the elementary equations (10.211) and (10.212) [7]. Then, the wave equation for the classical field $\langle E(r_\perp, t) \rangle$ is equivalently obtained as singlet approximation of this operator wave equation. To obtain the wave equation in the form of Eq. (10.247), we assumed that the light field propagates in the direction $r_\perp$ perpendicular to a planar structure with the background refractive index $n(r_\perp)$.

Equations (10.241), (10.242), (10.243) and (10.247) constitute the *Maxwell-semiconductor Bloch equations* (MSBE) [4, 47] which can be used as a general starting point to investigate excitations induced by classical light fields in direct-gap semiconductors. The form of the MSBE is formally exact and the quality of the results depends only on how accurately $\Gamma^{\lambda,\lambda'}$ and Γ^{QED} can be evaluated.

In the remainder of this section, we concentrate on situations where the Coulomb and classical light-induced effects dominate and the photon-assisted correlation terms can be omitted. In particular, we are interested to see which physical effects can be described via the MSBE by including $\Gamma^{\lambda,\lambda'}$ at different levels.

10.5.2 Excitonic States

The homogeneous solution of Eq. (10.241) without the two-particle correlations defines the eigenvalue problem known as the *Wannier equation*:

$$\tilde{\epsilon}_{\mathbf{k}_\parallel}\varphi^{\mathrm{R}}_{\lambda}(\mathbf{k}_\parallel) - \left(1 - f^{\mathrm{e}}_{\mathbf{k}_\parallel} - f^{\mathrm{h}}_{\mathbf{k}_\parallel}\right)\sum_{\mathbf{k}'_\parallel} V_{\mathbf{k}_\parallel-\mathbf{k}'_\parallel}\varphi^{\mathrm{R}}_{\lambda}(\mathbf{k}'_\parallel) = E_\lambda \varphi^{\mathrm{R}}_{\lambda}(\mathbf{k}_\parallel) \tag{10.249}$$

which, for vanishing densities, has a one-to-one correspondence to the Schrödinger equation for the relative-motion problem of atomic hydrogen [4]. The solutions of Eq. (10.249) define the *exciton* states which describe how electrons and holes are bound together due to the attractive Coulomb interaction of these oppositely charged quasi-particles.

As soon as carrier populations are present, $f^{\mathrm{e}}_{\mathbf{k}_\parallel}$ and $f^{\mathrm{h}}_{\mathbf{k}_\parallel}$ assume finite values with the consequence that Eq. (10.249) deviates from the original hydrogen problem and becomes a non-hermitian equation. Consequently, Eq. (10.249) has both left-handed, $\varphi^{\mathrm{L}}_{\lambda}(\mathbf{k}_\parallel)$, and right-handed, $\varphi^{\mathrm{R}}_{\lambda}(\mathbf{k}_\parallel)$, solutions connected via

$$\varphi^{\mathrm{L}}_{\lambda}(\mathbf{k}_\parallel) = \frac{\varphi^{\mathrm{R}}_{\lambda}(\mathbf{k}_\parallel)}{1 - f^{\mathrm{e}}_{\mathbf{k}_\parallel} - f^{\mathrm{h}}_{\mathbf{k}_\parallel}}. \tag{10.250}$$

These left- and right-handed solutions are normalized such that

$$\sum_{\mathbf{k}_\parallel}\varphi^{\mathrm{L}}_{\lambda}(\mathbf{k}_\parallel)\varphi^{\mathrm{R}}_{\nu}(\mathbf{k}_\parallel) = \delta_{\lambda,\nu}. \tag{10.251}$$

Since Eq. (10.249) defines a real-valued eigenvalue problem, we may choose the eigenstates to be real valued in momentum space.

When we take another look at the term

$$\left(1 - f^{\mathrm{e}}_{\mathbf{k}_\parallel} - f^{\mathrm{h}}_{\mathbf{k}_\parallel}\right) V_{\mathbf{k}_\parallel-\mathbf{k}'_\parallel} \equiv V^{\mathrm{eff}}_{\mathbf{k}_\parallel-\mathbf{k}'_\parallel}, \tag{10.252}$$

we notice that it can be viewed as an effective interaction. Since the phase-space filling factor, $(1 - f^{\mathrm{e}} - f^{\mathrm{h}})$, becomes negative for elevated densities, V^{eff} changes its sign for a certain range of momentum values as the density is increased. Consequently, the effective Coulomb interaction changes from attractive for low densities to repulsive for sufficiently large densities due to the Fermionic Pauli-blocking effects. The resulting Fermi pressure prevents the existence of bound excitons in the many-body system, i.e., the excitonic Mott transition is then reached [4, 24, 48, 49, 50].

In order to simplify the polarization equation, we use the excitonic states as a basis and expand the polarization:

$$P_{\mathbf{k}_{\|}} = \sum_{\lambda} p_{\lambda} \varphi_{\lambda}^{\mathrm{R}}(\mathbf{k}_{\|}), \qquad p_{\lambda} = \sum_{\mathbf{k}_{\|}} \varphi_{\lambda}^{\mathrm{L}}(\mathbf{k}_{\|}) P_{\mathbf{k}_{\|}}. \tag{10.253}$$

This way, Eq. (10.241) can be rewritten as

$$\mathrm{i}\hbar \frac{\partial}{\partial t} p_{\lambda} = E_{\lambda} p_{\lambda} - d_{vc} \sqrt{\mathsf{S}} \varphi_{\lambda}^{\mathrm{R}}(\mathbf{r}_{\|} = 0) \langle E(t) \rangle - \mathrm{i}\Gamma_{\lambda}. \tag{10.254}$$

10.5.3 Pump-Probe Calculations

Conceptually, the simplest experiment is to probe the linear response of the excited semiconductor with a weak classical probe spectrally overlapping the interesting transitions in the vicinity of the band-gap energy. The basic measurable quantities in this setup follow from the linear susceptibility,

$$\chi(\omega) \equiv \frac{P(\omega)}{\varepsilon_0 E(\omega)}, \tag{10.255}$$

which is obtained as the probe-induced macroscopic polarization, $P(\omega)$, divided by the probe field, $E(\omega)$. The imaginary part of the susceptibility is directly related to the semiconductor absorption.

Before we discuss the full microscopic correlation contributions to the SBE, we first summarize the analytic solution of the linear problem. As a simplification, we use a phenomenological expression for Γ [4] since this allows us to identify the principal effects beyond the coherent limit. To simplify the analysis, we start from an incoherent semiconductor system, i.e., all polarizations vanish before the system is excited. In this linear limit, $f_{\mathbf{k}_{\|}}^{\mathrm{e}}$ and $f_{\mathbf{k}_{\|}}^{\mathrm{h}}$ remain zero while only a small – linear – polarization $P_{\mathbf{k}_{\|}}$ is generated. When the microscopic Γ is replaced by a phenomenological value $-\mathrm{i}\gamma p_{\lambda}$, Eq. (10.254) becomes

$$\hbar\omega p_{\lambda}(\omega) = (E_{\lambda} - \mathrm{i}\gamma) P_{\lambda}(\omega) - d_{vc} \sqrt{\mathsf{S}}_{\varphi_{\lambda}^{\mathrm{R}}}(r = 0) \langle E(\omega) \rangle, \tag{10.256}$$

where we Fourier transformed to the frequency space. The solution of Eq. (10.256) determines the macroscopic polarization according to

$$P(\omega) = \frac{d_{c,v}}{\mathsf{S}} \sum_{\mathbf{k}_\parallel} P_{\mathbf{k}_\parallel}(\omega) = \frac{d_{c,v}}{\mathsf{S}} \sum_{\lambda} \sum_{\mathbf{k}_\parallel} \varphi_\lambda^{\mathrm{R}}(\mathbf{k}_\parallel) p_\lambda(\omega) =$$

$$= \frac{d_{c,v}}{\sqrt{\mathsf{S}}} \sum_{\lambda} \varphi_\lambda^{\mathrm{R}}(\mathbf{r}_\parallel = 0) p_\lambda(\omega) = |d_{cv}|^2 \sum_{\lambda} \frac{\left|\varphi_\lambda^{\mathrm{R}}(\mathbf{r}_\parallel = 0)\right|^2}{E_\lambda - \hbar\omega - \mathrm{i}\gamma} \langle E(\omega) \rangle, \qquad (10.257)$$

where the last step follows from the solution of Eq. (10.256). Inserting Eq. (10.257) into Eq. (10.255), we find the famous *Elliott formula* [51] for the linear semiconductor susceptibility

$$\chi(\omega) = \frac{|d_{cv}|^2}{\varepsilon_0} \sum_{\lambda} \frac{|\varphi_\lambda^{\mathrm{R}}(r = 0)|^2}{E_\lambda - \hbar\omega - \mathrm{i}\gamma}. \qquad (10.258)$$

Since the linear absorption is basically proportional to the imaginary part of the susceptibility [4], the semiconductor absorption shows resonances at the frequencies $\omega = E_\lambda/\hbar$ corresponding to excitonic energies.

The presence of excitonic resonances in $\chi(\omega)$ should not be taken as evidence for the existence of exciton populations in the probed systems. In fact, the largest and best defined resonances are observed for an originally unexcited low-temperature semiconductor. In this case, the linear response does not involve any populations and the weak probe field merely tests the transition possibilities of the interacting system. In other words, the linear response is exclusively determined by the linear polarization that defines the strengths of the different allowed optical transitions. When the linear response shows well defined, pronounced excitonic resonance, this only means that the light–matter-coupling-induced transitions are particularly strong at these frequencies.

To illustrate the basic features of the linear optical properties, we present in Fig. 10.4 the computed $\mathrm{Im}[\chi(\omega)]$ for an unexcited quantum-wire (QWI) and quantum-well (QW) system. We assumed phenomenological dephasing constants for which we took the values $\gamma = 0.19$ meV and $\gamma = 0.38$ meV. Comparing the QW and the QWI results, we immediately notice that the spectra look very similar. In both cases, they are dominated by excitonic resonances whose spectral width is determined by the phenomenological dephasing constant. From the energetically higher excitons, only the $n = 2$ state is well resolved. The other resonances merge with the onset of the continuum absorption. A more pronounced difference between the QW and QWI systems is visible in the results where we switched off the Coulomb interaction. Here, we obtain a peak just above the band-gap energy (12 meV above the 1 s resonance) for the QWI case as a consequence of the broadened $1/\sqrt{\hbar\omega}$ singularity of the 1D density of states. For the QW system, the density of states is a step function.

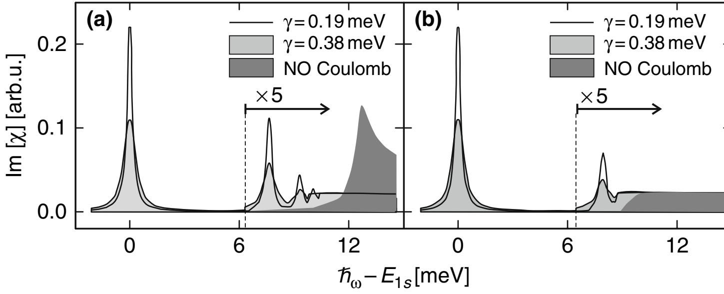


Fig. 10.4 Imaginary part of the susceptibility, $\mathrm{Im}[\chi(\omega)]$, obtained by evaluating the Elliott formula with a constant dephasing . Results for (**a**) a quantum-wire (QWI) system and (**b**) a quantum well (QW) are shown; to enhance the visibility of higher excitonic resonances, the corresponding spectrum is multiplied by 5. The calculated spectra for $\gamma = 0.19$ meV are plotted as a *solid line*, whereas the *light shaded area* presents the results for $\gamma = 0.38$ meV. For comparison, we plot as a *dark shaded area* the spectra obtained from a calculation without Coulomb interaction ($\gamma = 0.38$ meV). The frequency detuning is chosen with respect to the lowest exciton resonance at the energy E_{1s}

10.5.4 Coherent and Incoherent Carrier Correlations

Our analysis shows that it is natural to separate the singlet contributions into coherent and incoherent parts, i.e., into the optical polarization and the carrier populations, respectively. As shown in the previous section, the coherently induced polarization $P_{\mathbf{k}_\parallel}$ decays on a picosecond timescale, whereas the characteristic lifetime of the incoherent densities $f^{\mathrm{e}}_{\mathbf{k}_\parallel}$ and $f^{\mathrm{h}}_{\mathbf{k}_\parallel}$ is in the range of several nanoseconds.

In the same way as the singlets, also the carrier doublets can be divided into coherent and incoherent correlations depending on their characteristic decay times. The character of a generic two-particle correlation can be deduced from its dynamics according to the free semiconductor contribution, Eq. (10.50). The free time evolution from this contribution is always of the form

$$\mathrm{i}\hbar\frac{\partial}{\partial t}c^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel}_{\lambda,\nu,\nu',\lambda'} = (\varepsilon_{\lambda'} + \varepsilon_{\nu'} - \varepsilon_{\nu} - \varepsilon_{\lambda})c^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel}_{\lambda,\nu,\nu',\lambda'}\,. \tag{10.259}$$

If the energy prefactor is of the order of the band gap or even twice the gap energy, then the complex-valued correlation is very sensitive on scattering and can dephase on a time scale as fast as the coherent polarization. It turns out that also all source terms of such a *coherent* correlation involves at least one coherent quantity such as the electric field or a microscopic polarization. Once, the light field and the polarization have decayed, a coherent correlation cannot be created anymore. With this analysis, we find that

$$\Delta\langle 2\rangle_{\mathrm{coh}} = \left\{ c^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel}_{c,c,c,v},\ c^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel}_{v,v,v,c},\ c^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel}_{v,v,c,c} \right\} \tag{10.260}$$

represent the coherent correlations, whereas all other correlations such as

$$\Delta\langle 2\rangle_{\text{inc}} = \left\{ c_{c,c,c,c}^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel},\ c_{v,v,v,v}^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel},\ c_{c,v,c,v}^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel} \right\} \tag{10.261}$$

are the *incoherent correlations.*

10.5.5 Linear Optical Polarization

Before we investigate the non-linear optical properties, we first study the response of a semiconductor to a weak optical excitation. Even though we allow for the presence of finite densities, we still include only contributions that are linear in the optical probe-induced polarization. Since we assume classical fields, we can omit the quantum-optical correction in Eq. (10.241), i.e.,

$$\mathrm{i}\hbar \frac{\partial}{\partial t} P_{\mathbf{k}_\parallel} = \tilde{\epsilon}_{\mathbf{k}_\parallel} P_{\mathbf{k}_\parallel} - \left[1 - f_{\mathbf{k}_\parallel}^{\mathrm{e}} - f_{\mathbf{k}_\parallel}^{\mathrm{h}}\right] \Omega_{\mathbf{k}_\parallel} + \Gamma_{\mathbf{k}_\parallel}^{v,c} \quad , \tag{10.262}$$

$$\Gamma_{\mathbf{k}_\parallel}^{v,c} \equiv \sum_{\nu,\mathbf{k}'_\parallel,\mathbf{q}_\parallel \neq \mathbf{0}} V_{\mathbf{q}_\parallel} \left[c_{v,\nu;\nu,c}^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel} + c_{v,c;c,c}^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel} - \left(c_{c,\nu;\nu,v}^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel} + c_{c,c;c,v}^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel} \right)^* \right]. \tag{10.263}$$

The dynamics of the carrier densities is at least quadratic in the field strength of the probe pulse. For the computation of the linear response, we therefore assume that they are not changed by the weak probe field.

To solve Eqs. (10.262) and (10.263), we have to evaluate the dynamics of the coherent carrier correlations $c_{v,\nu;\nu,c}$, $c_{v,c;c,c}$, $c_{c,\nu;\nu,v}$, and $c_{c,c;c,v}$. Since the formal dynamics of all of these terms is very similar, we only give the results for $c_{v,\nu;\nu,c}$. Furthermore, we elaborate here only those parts of $c_{v,\nu;\nu,c}$ that are important for the linear response.

In general, $c_{v,\nu;\nu,c}$ couples also to incoherent density–density correlations $c_{\lambda,\lambda;\lambda,\lambda}$ and to correlations of the form

$$c_X^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel} \equiv c_{c,v;c,v}^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel} \tag{10.264}$$

which can be related to correlations of true exciton populations [17, 21]. Also c_X and $c_{\lambda,\lambda;\lambda,\lambda}$ are driven by the coherent light. However, these contributions are non-linear such that c_X and $c_{\lambda,\lambda;\lambda,\lambda}$, if it exists, only contribute as constant source to the linear response. In addition to these, $c_{v,\nu;\nu,c}$ also couples to the coherent biexciton amplitude

$$c_{\mathrm{BiX}}^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel} \equiv c_{v,v;c,c}^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel} \tag{10.265}$$

which is a coherent correlation. However, $c_{\rm BiX}$ is irrelevant for the linear response since it produces only non-linear contributions to the dynamics of $c_{v,v;v,c}$.

By using the elementary operator equations, we may calculate the dynamics of $c_{v,v;v,c}$ [1]. We only give the result including only those terms that are relevant for the linear response to a classical field,

$$\mathrm{i}\hbar\frac{\partial}{\partial t}c_{v,v;v,c}^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel} = \left(\tilde{\epsilon}^{\rm e}_{\mathbf{k}_\parallel-\mathbf{q}_\parallel} + \tilde{\epsilon}^{\rm h}_{\mathbf{k}_\parallel} - \tilde{\epsilon}^{\rm h}_{\mathbf{k}'_\parallel+\mathbf{q}_\parallel} + \tilde{\epsilon}^{\rm h}_{\mathbf{k}'_\parallel} - \mathrm{i}\gamma\right)c_{v,v;v,c}^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel} + \mathrm{S}_{v,v;v,c}^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel} + \left[\mathrm{D}_{v,v;v,c}^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel}\right]_{\rm coh} + \left[\mathrm{D}_{v,v;v,c}^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel}\right]_{\rm inc}, \tag{10.266}$$

where the triplet scattering is replaced by a phenomenological dephasing constant γ.

In general, $\mathrm{S}_{v,v;v,c}$ results from the singlet factorization of the Coulomb-induced three-particle terms. Physically, $\mathrm{S}_{v,v;v,c}$ acts as a source that generates $c_{v,v;v,c}$ even when the doublet correlations initially vanish. Explicitly, we can write

$$\mathrm{S}_{\lambda,\nu;\nu',\lambda'}^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel} \equiv \delta_{\sigma,\sigma'}V_{\mathbf{j}_\parallel}\left[P_{\mathbf{k}_\parallel-\mathbf{q}_\parallel}\left(f^{\rm h}_{\mathbf{k}_\parallel}f^{\rm h}_{\mathbf{k}'_\parallel}\bar{f}^{\rm h}_{\mathbf{k}'_\parallel+\mathbf{q}_\parallel}\right)_\Sigma - P_{\mathbf{k}'_\parallel}\left(f^{\rm h}_{\mathbf{k}_\parallel}f^{\rm e}_{\mathbf{k}_\parallel-\mathbf{q}_\parallel}\bar{f}^{\rm h}_{\mathbf{k}'_\parallel+\mathbf{q}_\parallel}\right)_\Sigma\right] + V_{\mathbf{q}_\parallel}\left[P_{\mathbf{k}_\parallel}\left(f^{\rm h}_{\mathbf{k}'_\parallel}f^{\rm e}_{\mathbf{k}_\parallel-\mathbf{q}_\parallel}\bar{f}^{\rm h}_{\mathbf{k}'_\parallel+\mathbf{q}_\parallel}\right)_\Sigma - P_{\mathbf{k}_\parallel-\mathbf{q}_\parallel}\left(f^{\rm h}_{\mathbf{k}_\parallel}f^{\rm h}_{\mathbf{k}'_\parallel}\bar{f}^{\rm h}_{\mathbf{k}'_\parallel+\mathbf{q}_\parallel}\right)_\Sigma\right], \tag{10.267}$$

where we have denoted the explicit spin-index dependency for the combination that is relevant for optical excitations, i.e.,

$$c_{v,v;v,c} \equiv c_{(v,\sigma),(v,\sigma');(v,\sigma'),(c,\sigma)}, \qquad S_{v,v;v,c} \equiv S_{(v,\sigma),(v,\sigma');(v,\sigma'),(c,\sigma)}. \tag{10.268}$$

We have also introduced the abbreviations

$$\bar{f}^\lambda_{\mathbf{k}_\parallel} = 1 - \bar{f}^\lambda_{\mathbf{k}_\parallel}, \tag{10.269}$$

$$\left(f^\lambda_{\mathbf{k}_\parallel}f^{\lambda'}_{\mathbf{k}'_\parallel}\bar{f}^{\lambda''}_{\mathbf{k}''_\parallel}\right)_\Sigma \equiv f^\lambda_{\mathbf{k}_\parallel}f^{\lambda'}_{\mathbf{k}'_\parallel}\left(1-f^{\lambda''}_{\mathbf{k}''_\parallel}\right) + \left(1-f^\lambda_{\mathbf{k}_\parallel}\right)\left(1-f^{\lambda'}_{\mathbf{k}'_\parallel}\right)\bar{f}^{\lambda''}_{\mathbf{k}''_\parallel}, \tag{10.270}$$

$$\mathbf{j}_\parallel \equiv \mathbf{k}'_\parallel + \mathbf{q}_\parallel - \mathbf{k}_\parallel. \tag{10.271}$$

Since $\mathrm{S}_{v,v;v,c}$ is the only term which drives the initially non-existing correlation, we conclude that $c_{v,v;v,c}$ is generated only via polarization transfer because all terms in $\mathrm{S}_{v,v;v,c}$ contain P. This observation also verifies that $c_{v,v;v,c}$ is a coherent correlation as classified earlier in Section 10.5.4.

Once $c_{v,v;v,c}$ is generated, it is modified by more complicated terms that contain the Coulomb-matrix element and correlated doublets in the singlet–doublet

factorization of the hierarchy problem. The explicit form of all terms relevant for the linear response is given by

$$\begin{aligned}
\left[D^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel}_{v,v;v,c}\right]_{\text{coh}} &= V_{\mathbf{q}_\parallel}\left(f^{\text{h}}_{\mathbf{k}'_\parallel+\mathbf{q}_\parallel}-f^{\text{h}}_{\mathbf{k}'_\parallel}\right)\sum_{\mathbf{l}_\parallel}\left(c^{\mathbf{q}_\parallel,\mathbf{l}_\parallel,\mathbf{k}_\parallel}_{v,c;c,c}+c^{\mathbf{q}_\parallel,\mathbf{l}_\parallel,\mathbf{k}_\parallel}_{v,v;v,c}\right)\\
&+V_{\mathbf{j}_\parallel}\left(f^{\text{h}}_{\mathbf{k}_\parallel}-f^{\text{h}}_{\mathbf{k}_\parallel+\mathbf{j}_\parallel}\right)\sum_{\mathbf{l}_\parallel}\left(c^{-\mathbf{j}_\parallel,\mathbf{k}'_\parallel,\mathbf{l}_\parallel}_{c,v;c,c}+c^{-\mathbf{j}_\parallel,\mathbf{k}'_\parallel,\mathbf{l}_\parallel}_{v,v;c,v}\right)\\
&+\left(1-f^{\text{h}}_{\mathbf{k}_\parallel}-f^{\text{h}}_{\mathbf{k}'_\parallel}\right)\sum_{\mathbf{l}_\parallel}V_{\mathbf{l}_\parallel+\mathbf{q}_\parallel}\left[c^{\mathbf{l}_\parallel,\mathbf{k}'_\parallel+\mathbf{q}_\parallel,\mathbf{k}_\parallel-\mathbf{q}_\parallel}_{c,v;v,v}\right]^*\\
&-\left(1-f^{\text{e}}_{\mathbf{k}_\parallel-\mathbf{q}_\parallel}-f^{\text{h}}_{\mathbf{k}_\parallel}\right)\sum_{\mathbf{l}_\parallel}V_{\mathbf{l}-\mathbf{k}_\parallel}c^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{l}_\parallel}_{v,v;v,c}\\
&+\left(1-f^{\text{e}}_{\mathbf{k}_\parallel-\mathbf{q}_\parallel}-f^{\text{h}}_{\mathbf{k}'_\parallel}\right)\sum_{\mathbf{l}_\parallel}V_{\mathbf{l}-\mathbf{k}'_\parallel}c^{\mathbf{j}_\parallel,\mathbf{l}_\parallel,\mathbf{k}_\parallel}_{v,v;v,c}\\
&+\left(f^{\text{h}}_{\mathbf{k}'_\parallel+\mathbf{q}_\parallel}-f^{\text{h}}_{\mathbf{k}_\parallel}\right)\sum_{\mathbf{l}_\parallel}V_{\mathbf{l}-\mathbf{k}_\parallel}c^{-\mathbf{j}_\parallel,\mathbf{k}'_\parallel,\mathbf{l}_\parallel}_{v,v;c,v}\\
&+\left(f^{\text{h}}_{\mathbf{k}'_\parallel}-f^{\text{h}}_{\mathbf{k}'_\parallel+\mathbf{q}_\parallel}\right)\sum_{\mathbf{l}_\parallel}V_{\mathbf{l}_\parallel-\mathbf{k}'_\parallel}c^{\mathbf{q}_\parallel,\mathbf{l}_\parallel,\mathbf{k}_\parallel}_{v,v;v,c}\\
&-\left(f^{\text{e}}_{\mathbf{k}_\parallel-\mathbf{q}_\parallel}-f^{\text{h}}_{\mathbf{k}'_\parallel+\mathbf{q}_\parallel}\right)\sum_{\mathbf{l}_\parallel}V_{\mathbf{l}_\parallel-\mathbf{q}_\parallel}c^{\mathbf{l}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel}_{v,v;v,c},
\end{aligned} \tag{10.272}$$

$$\begin{aligned}
\left[D^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel}_{v,v;v,c}\right]_{\text{inc}} &= V_{\mathbf{q}_\parallel}\left(P_{\mathbf{k}_\parallel}-P_{\mathbf{k}_\parallel-\mathbf{q}_\parallel}\right)\sum_{\mathbf{l}_\parallel}\left(c^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{l}_\parallel}_{c,v;v,c}+c^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{l}_\parallel}_{v,v;v,v}\right)\\
&-V_{\mathbf{j}_\parallel}\left(P_{\mathbf{k}'_\parallel}-P_{\mathbf{k}'_\parallel-\mathbf{j}_\parallel}\right)\sum_{\mathbf{l}_\parallel}\left(c^{-\mathbf{j}_\parallel,\mathbf{l}_\parallel,\mathbf{k}_\parallel}_{v,c;c,v}+c^{-\mathbf{j}_\parallel,\mathbf{l}_\parallel,\mathbf{k}_\parallel}_{v,v;v,v}\right)\\
&+\sum_{\mathbf{l}_\parallel}V_{\mathbf{l}-\mathbf{k}_\parallel}\left(P_{\mathbf{k}_\parallel-\mathbf{q}_\parallel}c^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{l}_\parallel}_{v,v;v,v}-P_{\mathbf{k}_\parallel}\left[c^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{l}_\parallel}_{c,v;v,c}-c^{-\mathbf{j}_\parallel,\mathbf{k}'_\parallel,\mathbf{l}_\parallel}_{c,v;c,v}\right]\right)\\
&-\sum_{\mathbf{l}_\parallel}V_{\mathbf{l}_\parallel-\mathbf{k}'_\parallel}\left(P_{\mathbf{k}'_\parallel}\left[c^{\mathbf{q}_\parallel,\mathbf{l}_\parallel,\mathbf{k}_\parallel}_{v,c;v,c}-c^{-\mathbf{j}_\parallel,\mathbf{l}_\parallel,\mathbf{k}_\parallel}_{v,c;c,v}\right]+P_{\mathbf{k}_\parallel-\mathbf{q}_\parallel}c^{\mathbf{j}_\parallel,\mathbf{k}_\parallel,\mathbf{l}_\parallel}_{v,v;v,v}\right)\\
&+\sum_{\mathbf{l}_\parallel}V_{\mathbf{l}_\parallel+\mathbf{q}_\parallel}\left[P^*_{\mathbf{k}_\parallel}c^{\mathbf{l}_\parallel,\mathbf{k}'_\parallel+\mathbf{q}_\parallel,\mathbf{k}_\parallel-\mathbf{q}_\parallel}_{c,v;v,c}+P^*_{\mathbf{k}'_\parallel}c^{\mathbf{l}_\parallel,\mathbf{k}'_\parallel+\mathbf{q}_\parallel,\mathbf{k}_\parallel-\mathbf{q}_\parallel}_{c,v;c,v}\right]^*\\
&-\sum_{\mathbf{l}_\parallel}V_{\mathbf{l}_\parallel-\mathbf{q}_\parallel}P_{\mathbf{k}_\parallel-\mathbf{q}_\parallel}c^{\mathbf{l}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel}_{v,v;v,v}.
\end{aligned} \tag{10.273}$$

The close inspection shows that the first two lines of Eqs. (10.272) and (10.273) contain terms where the Coulomb-matrix element appears outside the sum. As discussed in [1], such contributions yield Lindhard-type screening to the polarization dynamics while the remaining, more complicated terms may lead to the formation of new quasi-particle correlations.

When we solve Eqs. (10.262), (10.263) and (10.266), (10.267), (10.268), (10.269), (10.270), (10.271), (10.272), (10.273), we obtain a fully microscopic description for the linear response to a classical probe. We have presented here only the dynamics of $c_{v,v;v,c}$. However, the other correlations can be obtained from Eqs. (10.266), (10.267), (10.268), (10.269), (10.270), (10.271), (10.272), (10.273) using the simple substitution rules

$$v \leftrightarrow c, \quad f^{\mathrm{e}} \rightarrow 1 - f^{\mathrm{h}}, \quad f^{\mathrm{h}} \rightarrow 1 - f^{\mathrm{e}} \tag{10.274}$$

and/or complex conjugation.

The incoherent quantities, such as f^{e}, f^{h}, $c_{c,v;c,v}$, $c_{v,v;v,v}$, and $c_{c,c;c,c}$, are not changed by the weak probe pulse such that they drive coherent correlations only as external sources defined by the excitation state of the semiconductor at the time when the system is probed. Hence, when one uses ultrafast probe pulses, the system of incoherent quasi-particle excitations can be regarded as quasi-stationary. Consequently, we take in the following f^{e}, f^{h}, $c_{c,v;c,v}$, $c_{v,v;v,v}$, and $c_{c,c;c,c}$ as stationary quantities determined by the incoherent excitation state of the system.

Since single-particle carrier densities are present when one probes an excited incoherent semiconductor many-body state, they always contribute to the Coulomb-induced scattering via Eq. (10.267). At the same time, there exists a large phase space of semiconductor states where the incoherent correlations are infinitesimally small even when large concentrations of incoherent quasi-particle excitations are present. For example, an uncorrelated electron–hole plasma produces only vanishingly small correlation contributions (10.273) to the scattering of polarization.

As the most prominent correlated state, the semiconductor may contain true excitons, i.e., the Coulomb-bound electron–hole pairs described by c_X. Such terms enter to the polarization dynamics exclusively via the $\left[D_{v,v;v,c}\right]_{\mathrm{inc}}$ term. Also different configurations of incoherent quantities can alter the non-radiative scattering and dephasing experienced by the optical polarization. Thus, the presence of carrier densities or incoherent correlations introduces *excitation-induced dephasing* to the coherences [4, 42, 52, 53, 54, 55]. In the following, we investigate this phenomenon for various carrier concentrations using the full microscopic theory.

10.5.6 Excitation-Induced Dephasing

The essence of excitation-induced dephasing can be understood as we investigate the linear response of a semiconductor under quasi-stationary plasma

conditions, i.e., we assume that both exciton populations and density–density correlations are negligibly small. In practice, this situation can be realized for elevated lattice temperatures and/or elevated carrier densities [56]. Under these conditions, $\left[D_{v,v;v,c}\right]_{\text{inc}}$ can be omitted from the analysis such that only the carrier densities determine the initial state of the probed semiconductor.

In simplified treatments, the full $c_{v,v;v,c}$ dynamics, Eq. (10.266), has been reduced further by omitting also $\left[D_{v,v;v,c}\right]_{\text{coh}}$. This can be justified if coherences live only for much shorter times than it takes to build up new quasi-particles via $\left[D_{v,v;v,c}\right]_{\text{coh}}$. In this case, the steady-state result of Eq. (10.266), without the coherent or incoherent $\left[D_{v,v;v,c}\right]$, produces the well-known second Born scattering approximation [4, 52] where one typically uses the steady-state solution of $c_{v,v;v,c}$ within the Markov limit. At this level, one obtains a computationally feasible scheme for semiconductor systems of any dimensionality. Since the second Born linear response results have already shown excellent agreement between theory and experiments for a wide range of parameters [4, 57, 58], it can be concluded that the underlying assumptions represent a good approximation to the conditions realized in the respective experiments.

For our numerical evaluations, we assume that we probe a system in which we have an incoherent electron–hole plasma with Fermi–Dirac quasi-equilibrium distributions of electrons and holes at the lattice temperature, $T = 40\,\text{K}$. Figure 5a and c, respectively, presents the computed absorption spectra for a QWI and a QW for three representative carrier densities. Figure 5b and d shows $(f^{\text{e}}_{\mathbf{k}_\parallel} + f^{\text{h}}_{\mathbf{k}_\parallel})$ to quantify the level of excitation. For the lowest density, the population factor $(f^{\text{e}}_{\mathbf{k}_\parallel} + f^{\text{h}}_{\mathbf{k}_\parallel})$ is way below unity. Consequently, in the corresponding spectra we observe clear absorption resonances at the 1 s and the 2 s energy. A closer look reveals that the 2 s resonance is spectrally broader than the 1 s peak. This shows one of the basic features of Coulomb interaction-induced dephasing, i.e., the higher excitonic states experience more dephasing than the lower ones. This trend is clearly opposite to that of pure radiative dephasing. We can estimate from Fig. 10.5 that the excitation-induced dephasing produces a broadening in the range of $\gamma = 1$ meV for the highest density used. Thus, even moderate densities already lead to dephasing rates which largely exceed the radiative decay $\Gamma^{\text{rad}}_{1\text{s},1\text{s}} = 20\mu\text{eV}$. Hence, for these conditions the self-consistent light–matter coupling effects become less prominent. As a general trend, we observe that the QW system experiences a bit larger excitation-induced dephasing than the QWI since the phase space for Coulomb scattering events is larger in two dimensions than in one.

For elevated densities, also the 1 s resonance is broadened and the absorption dip between the bound and continuum states is gradually filled. This absorption increase is not a consequence of the band-gap shift but is caused by the excitation-induced resonance broadening, i.e., the frequency-dependent scattering [59]. Even for situations where $(f^{\text{e}}_{\mathbf{k}_\parallel} + f^{\text{h}}_{\mathbf{k}_\parallel})$ is still relatively low, the 2 s and higher excitons are already bleached. As a general feature for both well and wire systems, we see that the spectral position of the 1 s resonance remains basically

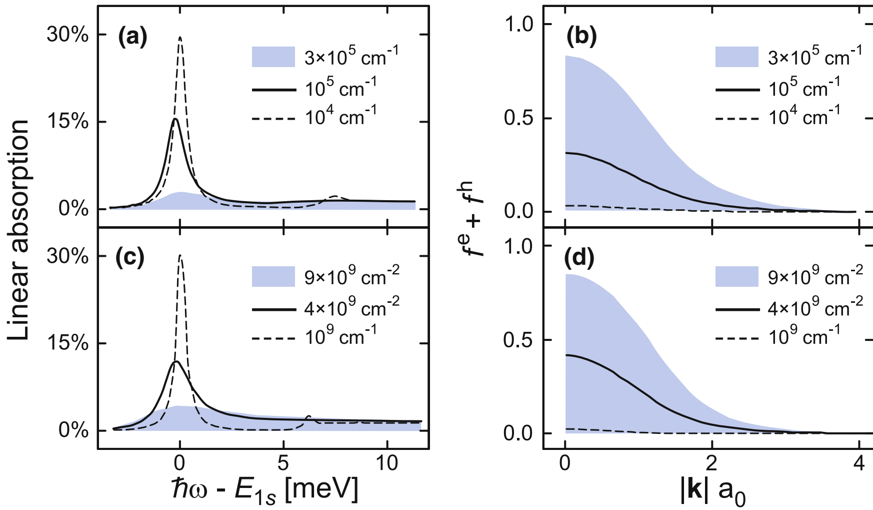


Fig. 10.5 Fully microscopically computed self-consistent absorption spectra. The QWI and QW spectra for three different carrier densities are shown in (**a**) and (**c**), respectively. Frames (**b**) and (**d**) show the assumed Fermi–Dirac quasi-equilibrium distributions $(f^{\mathrm{e}} + f^{\mathrm{h}})$ for the temperature 40 K

unchanged for different carrier densities, indicating that the microscopic scattering leads to energy renormalizations which compensate the Hartree–Fock shifts. As the density is increased, we see that the 1 s resonance is nearly completely bleached. The corresponding $(f^{\mathrm{e}}_{\mathbf{k}_\parallel} + f^{\mathrm{h}}_{\mathbf{k}_\parallel})$ is close to unity, indicating strong phase-space filling effects which eventually eliminate the bound exciton states. Only ionized excitons exist beyond this Mott transition [24, 48, 60]. As the density is increased further, the system enters to the regime of negative absorption, i.e., optical gain [61, 62, 63, 64].

Our numerical evaluations show that the full QWI computation and the second Born results are very similar for the investigated conditions. Hence, we conclude that the contributions of $\left[D_{v,v;v,c}\right]_{\mathrm{coh}}$ are not significant for the plasma conditions analyzed here. However, in cases where quasi-particle correlations are present, both the $\left[D_{v,v;v,c}\right]_{\mathrm{coh}}$ and $\left[D_{v,v;v,c}\right]_{\mathrm{inc}}$ contributions become important.

10.6 Semiconductor Quantum Optics

In this section, we apply our theory to treat quantum-optical effects such as photoluminescence, squeezing, and entanglement [7, 65, 66, 67, 68, 69]. The same theory can also be applied to describe quantum-optical effects in THz emission [70, 71, 72, 73]. In general, the quantum aspects of light

become particularly important when the light and carrier systems enter the incoherent regime. For these situations, the energy of the light field is completely stored in its quantum fluctuations such that all light–matter coupling effects are determined by quantum-optical aspects. Thus, both the light field and the related quasi-particle excitations must be treated fully quantum mechanically. As examples, we analyze luminescence and quantum-optical spectroscopy focusing on the differences in the semiconductor quasi-particle states resulting from quantum-optical instead of classical excitation.

10.6.1 Semiconductor Luminescence Equations

As a first step, we discuss how the light quantization effects the carrier system. Since we assume the completely incoherent regime, all coherent quantities (see discussion in Section 10.5.4) vanish. Consequently, the carrier densities are the only relevant single-particle variables.

Equation (10.242) from the previous section already contained the additional coupling to the quantum-optical correlations. Restricting ourselves to the incoherent regime, we find

$$\begin{aligned}\frac{\partial}{\partial t} f^{\mathrm{e}}_{\mathbf{k}_\parallel}\bigg|_{\mathrm{QED}} &= \frac{\mathrm{i}}{\hbar}\sum_{\mathbf{q}_\parallel}\left[\Delta\left\langle E^{v}_{\mathbf{q}_\parallel} a^{\dagger}_{c,\mathbf{k}_\parallel} a_{v,\mathbf{k}_\parallel-\mathbf{q}_\parallel}\right\rangle - \Delta\left\langle \left(E^{v}_{\mathbf{q}_\parallel}\right)^{\dagger} a^{\dagger}_{v,\mathbf{k}_\parallel-\mathbf{q}_\parallel} a_{c,\mathbf{k}_\parallel}\right\rangle\right] \\ &= -2\mathrm{Re}\left[\sum_{\mathbf{q}_\parallel,q_\perp} \mathsf{F}^{v,*}_{\mathbf{q}_\parallel,q_\perp}\Delta\left\langle B^{\dagger}_{\mathbf{q}_\parallel,q_\perp} a^{\dagger}_{v,\mathbf{k}_\parallel-\mathbf{q}_\parallel} a_{c,\mathbf{k}_\parallel}\right\rangle\right],\end{aligned} \tag{10.275}$$

where we have explicitly expressed Eq. (10.246) for $\Gamma^{\mathrm{QED}}_{c,c}$ and put the coherent correlations $\Delta\langle B a^{\dagger}_{v} a_{c}\rangle$ and $\Delta\langle B^{\dagger} a^{\dagger}_{c} a_{v}\rangle$ to zero. Similarly, we find the equation for the hole distributions:

$$\frac{\partial}{\partial t} f^{\mathrm{h}}_{\mathbf{k}_\parallel}\bigg|_{\mathrm{QED}} = -2\mathrm{Re}\left[\sum_{\mathbf{q}_\parallel,q_\perp} \mathsf{F}^{v,*}_{\mathbf{q}_\parallel,q_\perp}\Delta\left\langle B^{\dagger}_{\mathbf{q}_\parallel,q_\perp} a^{\dagger}_{v,\mathbf{k}_\parallel} a_{c,\mathbf{k}_\parallel-\mathbf{q}_\parallel}\right\rangle\right]. \tag{10.276}$$

From these equations we notice that the electron and hole densities couple to correlated processes, $\Delta\langle B^{\dagger} a^{\dagger}_{v} a_{c}\rangle$, where a photon is created by annihilating an electron–hole pair. The term $\Delta\langle B^{\dagger} a^{\dagger}_{v} a_{c}\rangle$ describes the correlated *photon-assisted electron–hole recombination*. In this process, the center-of-mass momentum $\mathbf{q}_\parallel$ of the electron–hole pair is conserved since the photon receives the same in-plane momentum. However, there is no momentum conservation in the $q_\perp$ direction for planar structures.

The expression $\Delta\langle B^\dagger a_v^\dagger a_c\rangle$ can also be interpreted as *photon-assisted polarization* because it contains the coupling of the photon to the polarization-type operator $a_v^\dagger a_c$. In the incoherent regime, this is the only relevant photon–carrier correlation and we define the abbreviation

$$\Pi_{\mathbf{k}_\|;\mathbf{q}_\|,q_\perp} \equiv \Pi^{v,c}_{\mathbf{k}_\|;\mathbf{q}_\|,q_\perp} = \Delta\left\langle B^\dagger_{\mathbf{q}_\|,q_\perp} a^\dagger_{v,\mathbf{k}_\|-\mathbf{q}_\|} a_{c,\mathbf{k}_\|}\right\rangle. \tag{10.277}$$

Because Eqs. (10.275) and (10.276) contain Π in a summed form, it is also convenient to identify the collective quantity

$$\Pi_{\mathbf{k}_\|;\mathbf{q}_\|,\Sigma} \equiv \sum_{q_\perp} \mathsf{F}^{v,*}_{\mathbf{q}_\|,q_\perp} \Pi_{\mathbf{k}_\|;\mathbf{q}_\|,q_\perp}. \tag{10.278}$$

that contains all Π terms having the same in-plane momentum, in analogy to Eq. (10.164).

The photon-assisted polarization terms appear also in the dynamics of the incoherent two-particle carrier correlations $c_{c,v;c,v}$, $c_{c,c;c,c}$, and $c_{v,v;v,v}$. Usually, the coupling of photons to exciton correlations, $c_X \equiv c_{c,v;c,v}$, introduces the largest effects. Therefore, we present here only the quantum-optical contributions to c_X:

$$\left.\frac{\partial}{\partial t} c_{\mathrm{X}}^{\mathbf{q}_\|,\mathbf{k}'_\|,\mathbf{k}_\|}\right|_{\mathrm{QED}} = -\left(1 - f^{\mathrm{e}}_{\mathbf{k}_\|} - f^{\mathrm{h}}_{\mathbf{k}_\|+\mathbf{q}_\|}\right)\Pi_{\mathbf{k}'_\|+\mathbf{q}_\|;\mathbf{q}_\|,\Sigma} - \left(1 - f^{\mathrm{e}}_{\mathbf{k}'_\|-\mathbf{q}_\|} - f^{\mathrm{h}}_{\mathbf{k}'_\|}\right)\Pi^{*}_{\mathbf{k}_\|-\mathbf{q}_\|;\mathbf{q}_\|,\Sigma}, \tag{10.279}$$

where all coherent contributions have again been omitted. Also the excitonic correlations are thus depleted by spontaneous emission.

From the definition of c_{X} in Eq. (10.264), it can be seen that $\mathbf{q}_\|$ plays the role of the exciton center-of-mass momentum. At the same time, $\mathbf{q}_\|$ appears in Eq. (10.279) as the in-plane photon momentum for the Π terms. Thus, the in-plane photon and exciton momenta have to match whenever photon-assisted processes either create or destroy photons or electron–hole pairs. Since the photon momentum is very small, the exciton correlations couple to the incoherent light field only when their center-of-mass momentum $\mathbf{q}_\|$ is nearly vanishing. This momentum selective coupling is important only for excitons and not for the carrier densities since their quantum dynamics in Eqs. (10.275) and (10.276) show that the carrier momentum $\mathbf{k}_\|$ can have any value and is not limited by the photon momentum.

In order to determine how the quantum-optical Π correlation influences the carrier dynamics, we have to calculate its Heisenberg equation of motion and apply the singlet–doublet factorization. The result in the incoherent limit is given by

$$
\begin{aligned}
\mathrm{i}\hbar\frac{\partial}{\partial t}\Pi_{\mathbf{k}_\parallel,\mathbf{q}_\parallel,q_\perp} = & \left(\tilde{\epsilon}^{c}_{\mathbf{k}_\parallel} - \tilde{\epsilon}^{v}_{\mathbf{k}_\parallel-\mathbf{q}_\parallel} - \hbar\omega_{\mathbf{k}_\parallel,\mathbf{q}_\parallel,q_\perp}\right)\Pi_{\mathbf{k}_\parallel,\mathbf{q}_\parallel,q_\perp} \\
& - \left[1 - f^{\mathrm{e}}_{\mathbf{k}_\parallel} - f^{\mathrm{h}}_{\mathbf{k}_\parallel-\mathbf{q}_\parallel}\right]\sum_{\mathbf{l}_\parallel} V_{\mathbf{k}_\parallel-\mathbf{l}_\parallel}\Pi_{\mathbf{l}_\parallel;\mathbf{q}_\parallel,q_\perp} \\
& + \mathrm{i}\hbar\mathsf{F}^{v}_{\mathbf{q}_\parallel,q_\perp}\left[f^{\mathrm{e}}_{\mathbf{k}_\parallel}f^{\mathrm{h}}_{\mathbf{k}_\parallel-\mathbf{q}_\parallel} + \sum_{\mathbf{l}_\parallel} c_X^{\mathbf{q}_\parallel,\mathbf{k}_\parallel-\mathbf{q}_\parallel,\mathbf{l}_\parallel}\right] \\
& - \mathrm{i}\hbar\left[1 - f^{\mathrm{e}}_{\mathbf{k}_\parallel} - f^{\mathrm{h}}_{\mathbf{k}_\parallel-\mathbf{q}_\parallel}\right]\Delta\left\langle B^\dagger_{\mathbf{q}_\parallel,q_\perp}B_{\mathbf{q}_\parallel,\Sigma}\right\rangle + T^{\Pi}_{\mathbf{k}_\parallel;\mathbf{q}_\parallel,q_\perp}\,, \qquad (10.280)
\end{aligned}
$$

where we have used the collective photon operator according to Eq. (10.164). The triplet term T^{Π} provides coupling to the three-particle correlations and is given by

$$
\begin{aligned}
\mathrm{T}^{\Pi}{}_{\mathbf{k}_\parallel;\mathbf{q}_\parallel,q_\perp} \equiv & V\left[\Pi^{v,c}_{\mathbf{k}_\parallel;\mathbf{q}_\parallel,q_\perp}\right]_T = \\
= & \sum_{\nu,\mathbf{k}'_\parallel,\mathbf{l}_\parallel}\left(V_{\mathbf{l}_\parallel}\Delta\left\langle B^\dagger_{\mathbf{q}_\parallel,q_\perp}a^\dagger_{v,\mathbf{k}_\parallel-\mathbf{q}_\parallel}a^\dagger_{\nu,\mathbf{k}'_\parallel+\mathbf{q}_\parallel}a_{\nu,\mathbf{k}'_\parallel+\mathbf{l}_\parallel}a_{c,\mathbf{k}_\parallel-\mathbf{l}_\parallel}\right\rangle\right. \\
& \left. - V_{\mathbf{l}_\parallel-\mathbf{q}_\parallel}\Delta\left\langle B^\dagger_{\mathbf{q}_\parallel,q_\perp}a^\dagger_{v,\mathbf{k}_\parallel-\mathbf{l}_\parallel}a^\dagger_{\nu,\mathbf{k}'_\parallel+\mathbf{l}_\parallel}a_{\nu,\mathbf{k}'_\parallel}a_{c,\mathbf{k}_\parallel}\right\rangle\right) \\
& - \mathrm{i}\hbar\sum_{\mathbf{l}_\parallel}\left(\Delta\left\langle B^\dagger_{\mathbf{q}_\parallel,q_\perp}B_{\mathbf{l}_\parallel,\Sigma}a^\dagger_{v,\mathbf{k}_\parallel-\mathbf{q}_\parallel}a_{v,\mathbf{k}_\parallel-\mathbf{l}_\parallel}\right\rangle\right. \\
& \left. - \Delta\left\langle B^\dagger_{\mathbf{q}_\parallel,q_\perp}B_{\mathbf{q}_\parallel-\mathbf{l}_\parallel,\Sigma}a^\dagger_{c,\mathbf{k}_\parallel-\mathbf{l}_\parallel}a_{c,\mathbf{k}_\parallel}\right\rangle\right) \\
& + \sum_{\mathbf{l}_\parallel}\left(\Delta\left\langle B^\dagger_{\mathbf{q}_\parallel,q_\perp}Q^{c}_{\mathbf{l}_\parallel}a^\dagger_{v,\mathbf{k}_\parallel-\mathbf{q}_\parallel}a_{c,\mathbf{k}_\parallel-\mathbf{l}_\parallel}\right\rangle\right. \\
& \left. - \Delta\left\langle B^\dagger_{\mathbf{q}_\parallel,q_\perp}Q^{v}_{\mathbf{q}_\parallel-\mathbf{l}_\parallel}a^\dagger_{v,\mathbf{k}_\parallel-\mathbf{l}_\parallel}a_{c,\mathbf{k}_\parallel}\right\rangle\right). \qquad (10.281)
\end{aligned}
$$

Here, the terms given by the first sum on the right hand side describe the influence of the Coulomb-induced scattering on the dynamics of Π, while the second and the third sums provide higher order correlations due to the coupling to photons and phonons. While the complete treatment of the three-particle level is far beyond current computer resources, one can use the analogy to the scattering terms of the coherent polarization and derive an approximate form at the scattering level as presented explicitly in Hoyer et al. [59] and Kira et al. [68].

In general, Eq. (10.280) shows that Π is spontaneously driven by the term $(f^{\mathrm{e}}f^{\mathrm{h}} + \sum c_X)$ in the third line even when all correlations initially vanish. Thus, this contribution acts as a *spontaneous emission source* which has a natural cluster-expansion-based division into its correlated c_X part and the $f^{\mathrm{e}}f^{\mathrm{h}}$ part related to the uncorrelated electron–hole plasma. We will see in Section 10.7

that only c_X can describe the effects of true exciton populations whereas the plasma contribution, $f^{\mathrm{e}}_{\mathbf{k}_\|} f^{\mathrm{h}}_{\mathbf{k}_\| - \mathbf{q}_\|}$, is an emission source due to the spontaneous recombination of uncorrelated electron–hole pairs. Clearly, this recombination process occurs as long as an electron and a hole are found simultaneously with momenta $\mathbf{k}_\|$ and $\mathbf{k}_\| - \mathbf{q}_\|$, respectively. The corresponding center-of-mass momentum $\mathbf{q}_\|$ is then transferred to the photon. The correlated c_X source can include contributions resulting from the presence of genuine exciton populations and/or a correlated electron–hole plasma [74]. Since neither the uncorrelated plasma nor the correlated sources depend on the photon frequency in any way, they both can initiate photon emission – that is observed as photoluminescence (PL) – in all relevant frequency ranges.

The spectral distribution of spontaneously emitted photons is strongly altered by the resonance structure related to the homogeneous part of the Π dynamics, i.e., the terms in the first and second line of Eq. (10.280). We see that the different $\mathbf{k}_\|$ components of the spontaneously generated Π are coupled via the Coulomb sum in the second line of Eq. (10.280). It is interesting to notice that this part of Eq. (10.280) shows strong analogies to the homogeneous part of the semiconductor Bloch equations. Hence, we see that it is the Coulomb coupling that produces the excitonic resonances in the resulting photoluminescence in the same way as these resonances appear in the absorption spectra. This important fact implies that the specific form of the quasi-particle state of carriers in the spontaneous emission source does not determine whether or not PL shows excitonic resonances. Thus, *the detection of an excitonic resonance in a luminescence spectrum cannot be a unique signature for the presence of exciton populations*, in contrast to resonances in the THz response. Instead, "excitonic luminescence" may also result from quasi-particle states containing only a pure electron–hole plasma. This intriguing phenomenon was first predicted [67] and later verified experimentally [56, 74, 75]. In general, the detailed signatures of plasma and exciton population contributions to the excitonic luminescence can be identified via a quantitative analysis [56, 59, 74].

The last line of Eq. (10.280) contains photon-number-like correlations that are particularly large when the semiconductor material either is inside an optical cavity or if it is optically pumped with incoherent light fields. Thus, this contribution provides either stimulated coupling or direct excitation effects due to external incoherent fields. The corresponding dynamics is described by

$$\mathrm{i}\hbar \frac{\partial}{\partial t} \Delta\left\langle B^\dagger_{\mathbf{q}_\|,q_\perp} B_{\mathbf{q}_\|,q'_\perp} \right\rangle = \hbar\left(\omega_{\mathbf{q}'} - \omega_{\mathbf{q}}\right) \Delta\left\langle B^\dagger_{\mathbf{q}_\|,q_\perp} B_{\mathbf{q}_\|,q'_\perp} \right\rangle + \mathrm{i}\hbar \sum_{\mathbf{k}_\|} \left[\mathsf{F}^{\nu}_{\mathbf{q}_\|,q_\perp} \Pi^*_{\mathbf{k}_\|;\mathbf{q}_\|,q'_\perp} + \mathsf{F}^{\nu,*}_{\mathbf{q}_\|,q'_\perp} \Pi_{\mathbf{k}_\|;\mathbf{q}_\|,q_\perp} \right], \quad (10.282)$$

where we again only included the incoherent correlations. If the carrier system is close to a quasi-equilibrium situation, the carrier quantities entering Eq. (10.280) are nearly constant. In this regime, Eqs. (10.280), (10.281) and (282)

are closed. They fully determine the photon flux for the emitted light providing the steady-state luminescence spectrum according to

$$I_{\mathrm{PL}}(\omega_{\mathbf{q}}) = \frac{\partial}{\partial t}\Delta\left\langle B^{\dagger}_{\mathbf{q}_{\|},q_{\perp}} B_{\mathbf{q}_{\|},q_{\perp}}\right\rangle = 2\mathrm{Re}\left[\sum_{\mathbf{k}_{\|}} \mathsf{F}^{*}_{\mathbf{q}_{\|},q_{\perp}} \Pi^{*}_{\mathbf{k}_{\|};\mathbf{q}_{\|},q_{\perp}}\right]. \tag{10.283}$$

In general, Eqs. (10.280), (10.281), and (10.282) define the *semiconductor luminescence equations* (SLE) [67, 76] since they have an obvious structural similarity to the semiconductor Bloch equations (SBE) discussed in Section 10.5.1. The SLE can be applied to systematically explain quantitative features of PL ranging from the low-density conditions [7, 77] up to the gain regime [63, 78, 79]. The SLE can also be generalized for excitations containing coherences. For these situations, also quantum-optical correlations of the type $\Delta\langle BB\rangle$, $\Delta\langle Ba^{\dagger}_{v}a_{c}\rangle$, and $\Delta\langle Ba^{\dagger}_{\lambda}a_{\lambda}\rangle$ become relevant. These contributions lead to new quantum-optical effects such as squeezing in the resonance fluorescence [68], entanglement-generated quantum oscillations [69], and resonances in the probe transmission [65]. For a review of the generalization of the theory toward the coherent regime, see Kira et al. [7].

10.6.2 *Radiative Recombination of Carriers and Exciton Populations*

To analyze the effect of spontaneous emission on the exciton and carrier distributions, Fig. 10.6 shows them at different time moments after coherent resonant 1s-excitation. These results have been obtained by solving the full singlet–doublet equations for a planar arrangement of quantum wires. In that case, the numerical complexity for the one-dimensional carrier correlations can be handled numerically, while a sufficiently high wire density is assumed such that the classical exciting light field can still be solved with the wave equation for planar situations.

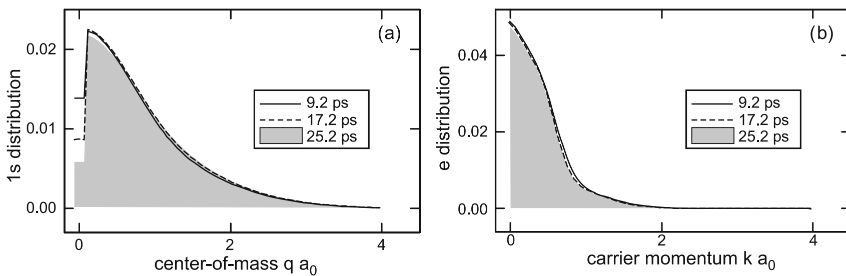


Fig. 10.6 (**a**) Computed 1s-exciton and (**b**) carrier distributions for a time sequence after the resonant 1s excitation with classical light. (After Kira and Koch [13])

We have assumed here a 1 ps excitation pulse with a sufficiently weak intensity such that an incoherent exciton fraction of more than 80% is obtained. The snapshot times in Fig. 10.6 are chosen such that the polarization-to-population conversion is already complete. This conversion is mostly due to acoustic phonon scattering which converts coherent polarization to incoherent excitons. Since the phonons transfer their momentum, a wide momentum spread can be observed not only for the carrier densities but also for the 1s-exciton distributions.

After the population generation, the excitons in the very low-momentum states, i.e., roughly those with $|\mathbf{q}_{\|}|a_0 < 0.1$, show a fast decay due to their photoluminescence-related recombination. In other words, these are the optically active *bright excitons* that give rise to luminescence. Due to its momentum selectivity this recombination can lead to a significant hole burning in the exciton distributions [17]. This hole burning is supported by the fact that the exciton scattering times are relatively slow in comparison to the relatively fast ≈15 ps recombination time, which is the same as that of the coherent polarization [1, 77]. Hence, the bright excitons are strongly coupled to the light field, which rapidly depletes their population leaving the majority of the excitons in optically inactive *dark states*.

Since an electron with an arbitrary momentum $\mathbf{k}_{\|}$ can recombine with a hole in the matching momentum state, $\mathbf{k}_{\|} - \mathbf{q}_{\|}$, all electron and hole states contribute to the emission, i.e., electron and hole distributions do not show a momentum selectivity. Consequently, the radiative recombination only leads to slow changes of their total electron–hole population on a nanosecond time scale. This important difference between exciton and carrier distributions is the reason for the fact that excitons display highly non-thermal distributions, even if the carriers are basically in a thermal quasi-equilibrium state. As a result, *spontaneous emission is never a weak perturbation for exciton distributions in the usual direct-gap semiconductors* even though the total carrier recombination rate is slow.

The discussed fundamental differences between the optical coupling of exciton and electron–hole populations lead to strong non-equilibrium features in the exciton photoluminescence [56, 59, 74, 80, 81, 82]. As an example, we show in Fig. 10.7 computed luminescence spectra, I_{PL}, for the different times used in Fig. 10.6. We observe that the luminescence decreases with increasing time,

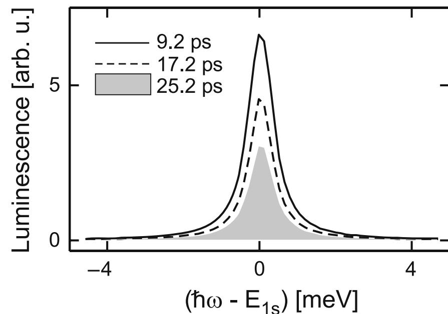


Fig. 10.7 Photoluminescence spectra computed for different times after the 1s excitation with classical field of Fig. 10.6. (After Kira and Koch [13])

following the depletion of the bright exciton states. However, rather strong excitonic PL remains, even after most of the optically active excitons have decayed. This *excitonic PL without exciton populations* underlines the fact that also the uncorrelated electron–hole plasma produces an excitonic resonance in the luminescence [56, 67, 75].

10.6.3 Correlated Photons in Quantum-Well Emission

The luminescence spectra investigated in the previous section offered a first important example for a true quantum-optical effect in semiconductors. More recently, however, also other quantum-optical effects known from atomic optics have been tested with semiconductor structures and the interesting combination between quantum optics and semiconductor physics has emerged. Semiconductor quantum dots are utilized to optically entangle excitonic states [83, 84] as well as to emit single photons [85], and light–matter entanglement is found to strongly influence semiconductor-cavity experiments [65, 69]. Here, we demonstrate how correlations between a photon and the complex many-body carrier system of a semiconductor quantum well can be visualized by interference patterns arising from a single photon spontaneously emitted into different directions. These patterns are realized by collecting light emitted from a quantum well into two different directions and to combine those light beams on a common detector.

Under steady-state conditions, the emission spectrum according to Eq. (10.283) is proportional to the rate of emitted photons which according to Eq. (10.282) can directly be evaluated from the photon-assisted polarizations. A closer investigation of Eq. (10.282) shows that in general also correlations of photons emitted into different directions can be formed, i.e., first-order photonic correlation functions of the form $\Delta\langle B^{\dagger}_{\mathbf{q}_{\|},+q_{\perp}} B_{\mathbf{q}_{\|},q'_{\perp}}\rangle$ with different values of $q'_{\perp}$. As we pointed out before, due to the lack of momentum conservation in the emission direction perpendicular to the QW system, no restrictions apply to the value of $q'_{\perp}$. Since for general emission directions the operator $\Delta\langle B^{\dagger}_{\mathbf{q}} B_{\mathbf{q}'}\rangle$ is not Hermitian, its expectation value is in general complex and not directly observable. Therefore, we investigate a setup sketched in Fig. 10.8. With the help of mirrors, photons emitted to both sides of the sample are redirected into a

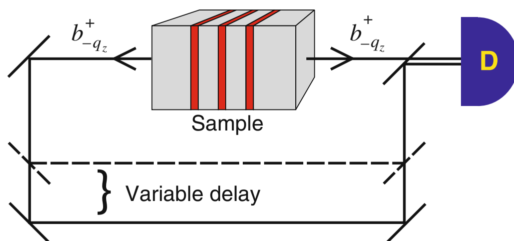


Fig. 10.8 Schematic setup of suggested interference measurement. Photons emitted at the two opposite sides of the sample are redirected into a common detector while the relative phase can be changed via a variable path length

common detector. One of the paths can be varied in its optical length such that the corresponding detector operator for normal emission is given by

$$d_{q_\perp} = \frac{1}{\sqrt{2}}\left(B_{\mathbf{q}_\parallel=0,q_\perp} + e^{\mathrm{i}\varphi}\, B_{\mathbf{q}_\parallel=0,-q_\perp}\right), \tag{10.284}$$

where φ depends on the difference in the optical path. The detected signal under steady-state conditions is again proportional to the photon flux, but now in the basis of detector operators,

$$\begin{aligned}\left\langle d^\dagger_{q_\perp} d_{q_\perp}\right\rangle &= \frac{1}{2}\left(\left\langle B^\dagger_{\mathbf{0},q_\perp} B_{\mathbf{0},q_\perp}\right\rangle + \left\langle B^\dagger_{\mathbf{0},-q_\perp} B_{\mathbf{0},-q_\perp}\right\rangle\right) \\ &\quad + \cos(\varphi)\left|\left\langle B^\dagger_{\mathbf{0},q_\perp} B_{\mathbf{0},-q_\perp}\right\rangle\right|. \end{aligned} \tag{10.285}$$

By varying φ, the cross-correlations can thus be determined. An example for the expected first-order cross-correlation

$$I_{\mathrm{CC}} \equiv \frac{\partial}{\partial t}\left|\left\langle B^\dagger_{\mathbf{q}_\parallel=0,+q_\perp} B_{\mathbf{q}_\parallel=0,-q_\perp}\right\rangle\right| \tag{10.286}$$

is shown in Fig. 10.9. There, we compare the usual luminescence spectrum with the result of the computed first-order cross-correlations between photons emitted to the two opposite directions perpendicular to the quantum well.

In order to understand the fundamental reason for the existence of cross-correlations, we investigate the particular form of light–matter interaction. By assuming carrier confinement to the lowest quantum-well subband, the dipole Hamiltonian, Eq. (10.162), contains operator combinations of the form $B^\dagger_{\mathbf{q}_\parallel,q_\perp} a^\dagger_{v,\mathbf{k}_\parallel} a_{c,\mathbf{k}_\parallel+\mathbf{q}_\parallel}$. This interaction allows processes where photons are emitted by simultaneous recombination of an electron–hole pair. As a consequence of the missing translational symmetry, the z-component of the momentum is not conserved. In the classical regime, this symmetry breaking is known to lead to radiative decay of quantum-well polarization [87, 88].

As a consequence of the non-conservation of momentum, the same state can symmetrically emit light to both left and right propagating modes. This leads to

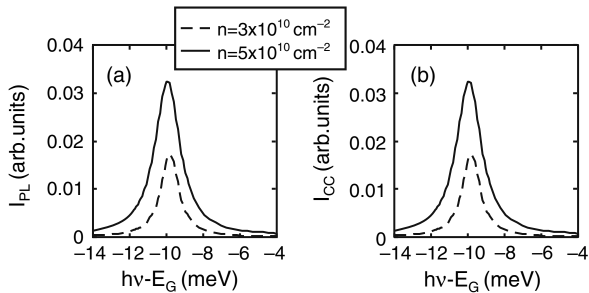


Fig. 10.9 Normalized photoluminescence spectrum of a single quantum well (**a**) compared to the spectrum of cross-correlations (**b**) for two different densities at a carrier temperature of 77 K. (After Hoyer et al. [86])

entanglement between $\pm q_{\perp}$ modes. Therefore, the suggested experiment is in close analogy to the double-slit experiment [89, 90, 91] performed with a single electron or photon. Even though the light emission is completely incoherent, the photon entanglement allows interference, i.e., the observation of cross-correlations, at the detector. The predicted correlations have been successfully observed as described in Hoyer et al. [66], and a more careful investigation of the angular resolved measurement is suggested to give a new quantum-optical method of spectroscopy for mapping out disorder landscapes [92].

10.7 Probing Incoherent Populations

In dilute gas spectroscopy, one often detects small concentrations of a particular species of atoms or molecules by using an optical probe that is sensitive to transitions between the eigenstates of the respective species. If the characteristic absorption resonances are observed in the probe spectrum, the atoms or molecules must be present, and one can deduce their relative concentration through proper normalization of the respective transition strength. To understand why this simple scenario does not apply for the detection of excitons via interband optics in semiconductors, we have to remember that the interband transitions in semiconductors do not conserve the number of electron–hole pairs. In other words, each interband absorption process creates an electron–hole pair while an interband emission process destroys such a pair. As a result, interband absorption or emission leads to transitions that connect semiconductor eigenstates with different numbers of electron–hole pairs.

In order to find a direct analogue between semiconductor optics and atomic spectroscopy, we have to consider an energy range of light that does not change the number of electron–hole-pair excitations, i.e., we need to consider intraband transitions where electron–hole pairs are neither created nor destroyed. In particular, we want to look for transitions between the excitonic levels [93, 94, 95, 96] to identify the presence of exciton quasi-particles in semiconductors. Here, the most pronounced resonance is expected at $\hbar\omega_{\mathrm{t}} = \hbar\omega_{2\mathrm{p}} - \hbar\omega_{1\mathrm{s}}$ corresponding to the excitation of the exciton from its lowest, 1 s, state to the next higher, 2 p, state. For many of the commonly studied direct-gap compound semiconductors, the excitonic binding energies are in the range of a few meV such that the transition energy $\hbar\omega_{\mathrm{t}}$ is in the terahertz (THz) part of the electromagnetic spectrum [93, 95].

In the following section, we discuss the direct correspondence between atomic spectroscopy and THz spectroscopy in semiconductors. A particular interest here is to find a direct way to detect the exciton number or more generally the presence of incoherent excitonic correlations. The theory for THz spectroscopy can be described microscopically with the same precision as the optical interband spectroscopy by applying the same cluster-expansion approach as we have used so far. However, here wedo not elaborate on the

details of the calculations and emphasize only the fact that THz spectroscopy can unambiguously identify true exciton populations. For this purpose, we only briefly summarize the main steps [14, 17, 74, 97, 98, 99] needed to understand linear THz absorption features.

10.7.1 Dynamics of Exciton Correlations

Before we investigate the THz response of the carrier system, we take a closer look at the excitonic correlation equation, i.e., at the dynamics of c_X. For this purpose, we write the singlet–doublet equations in the form

$$\begin{aligned} i\hbar\frac{\partial}{\partial t}c_X^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel} &= \left(\tilde{\epsilon}^e_{\mathbf{k}'_\parallel+\mathbf{q}_\parallel} + \tilde{\epsilon}^h_{\mathbf{k}'_\parallel} - \tilde{\epsilon}^e_{\mathbf{k}_\parallel} - \tilde{\epsilon}^h_{\mathbf{k}_\parallel-\mathbf{q}_\parallel}\right)c_X^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel} + S_X^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel} \\ &+ \left(1 - f^{\mathrm{e}}_{\mathbf{k}_\parallel} - f^{\mathrm{h}}_{\mathbf{k}_\parallel-\mathbf{q}_\parallel}\right)\sum_{\mathbf{l}_\parallel} V_{\mathbf{l}_\parallel-\mathbf{k}_\parallel} c_{\mathrm{X}}^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{l}_\parallel} \\ &- \left(1 - f^{\mathrm{e}}_{\mathbf{k}'_\parallel+\mathbf{q}_\parallel} - f^{\mathrm{h}}_{\mathbf{k}'_\parallel}\right)\sum_{\mathbf{l}_\parallel} V_{\mathbf{l}_\parallel-\mathbf{k}'_\parallel} c_{\mathrm{X}}^{\mathbf{q}_\parallel,\mathbf{l}_\parallel,\mathbf{k}_\parallel} \\ &+ G_{X,\mathrm{Coul}}^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel} + G_{X,\mathrm{phon}}^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel} + D_{X,\mathrm{rest}}^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel} + T_X^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel}. \end{aligned} \tag{10.287}$$

Here, the first line is the sum of the renormalized kinetic energy of the particles plus the singlet source:

$$\begin{aligned} \mathrm{S}_X^{\mathbf{q}_\parallel,\mathbf{k}'_\parallel,\mathbf{k}_\parallel} \equiv \delta_{\sigma,\sigma'} V_{\mathbf{j}_\parallel}\Big[&\left(f^{\mathrm{e}}_{\mathbf{k}'_\parallel+\mathbf{q}_\parallel} f^{\mathrm{h}}_{\mathbf{k}'_\parallel} \bar{f}^e_{\mathbf{k}_\parallel} \bar{f}^h_{\mathbf{k}_\parallel-\mathbf{q}_\parallel}\right)_\Sigma \\ &+ P^*_{\mathbf{k}_\parallel} P_{\mathbf{k}'_\parallel+\mathbf{q}_\parallel}(f^{\mathrm{h}}_{\mathbf{k}_\parallel-\mathbf{q}_\parallel} - f^{\mathrm{h}}_{\mathbf{k}'_\parallel}) + P^*_{\mathbf{k}_\parallel-\mathbf{q}_\parallel} P_{\mathbf{k}'_\parallel}\left(f^e_{\mathbf{k}_\parallel} - f^{\mathrm{e}}_{\mathbf{k}'_\parallel+\mathbf{q}_\parallel}\right)\Big] \\ &+ V_{\mathbf{q}_\parallel}\Big[P^*_{\mathbf{k}_\parallel} P_{\mathbf{k}'_\parallel}\left(f^{\mathrm{h}}_{\mathbf{k}_\parallel-\mathbf{q}_\parallel} - f^{\mathrm{e}}_{\mathbf{k}'_\parallel+\mathbf{q}_\parallel}\right) - P^*_{\mathbf{k}_\parallel-\mathbf{q}_\parallel} P_{\mathbf{k}'_\parallel+\mathbf{q}_\parallel}\left(f^{\mathrm{h}}_{\mathbf{k}'_\parallel} - f^{\mathrm{e}}_{\mathbf{k}_\parallel}\right) \\ &- P^*_{\mathbf{k}_\parallel-\mathbf{q}_\parallel} P_{\mathbf{k}'_\parallel}\left(f^{\mathrm{e}}_{\mathbf{k}_\parallel} - f^{\mathrm{e}}_{\mathbf{k}'_\parallel+\mathbf{q}_\parallel}\right) + P^*_{\mathbf{k}_\parallel} P_{\mathbf{k}'_\parallel+\mathbf{q}_\parallel}\left(f^{\mathrm{h}}_{\mathbf{k}'_\parallel} - f^{\mathrm{h}}_{\mathbf{k}_\parallel-\mathbf{q}_\parallel}\right)\Big]. \end{aligned} \tag{10.288}$$

This expression contains the singlet factorization of the Coulomb-induced two- and three-particle terms. For clarity, we explicitly write here the spin dependence following from the sequence $c_X \equiv c_{(c,\sigma),(v,\sigma');(c,\sigma'),(v,\sigma)}$. Additionally, we introduce the abbreviation

$$\begin{aligned} \left(f^\lambda_{\mathbf{k}_\parallel} f^{\lambda'}_{\mathbf{k}'_\parallel} \bar{f}^{\lambda''}_{\mathbf{k}''_\parallel} \bar{f}^{\lambda'''}_{\mathbf{k}'''_\parallel}\right)_\Sigma &\equiv f^\lambda_{\mathbf{k}_\parallel} f^{\lambda'}_{\mathbf{k}'_\parallel}\left(1 - f^{\lambda''}_{\mathbf{k}''_\parallel}\right)\left(1 - f^{\lambda'''}_{\mathbf{k}'''_\parallel}\right) \\ &- \left(1 - f^\lambda_{\mathbf{k}_\parallel}\right)\left(1 - f^{\lambda'}_{\mathbf{k}'_\parallel}\right) f^{\lambda''}_{\mathbf{k}''_\parallel} f^{\lambda'''}_{\mathbf{k}'''_\parallel}, \end{aligned} \tag{10.289}$$

which identifies the in- and out-scattering terms similar to the second Born scattering source. These terms act as a source to the c_X dynamics also in the purely incoherent regime, verifying once again that c_X is fundamentally an incoherent correlation.

The second and third lines of Eq. (10.287) contain the two most important contributions of the incoherent Coulomb-induced correlations $[D_X]_{\mathrm{inc}}$. In general, $[D_X]_{\mathrm{inc}}$ consists of terms such as $(n^{\lambda} - n^{\nu}) \sum V c_{\lambda,\nu;\nu',\lambda'}$. As a result, $[D_X]_{\mathrm{inc}}$ can be considered as a systematic generalization of the single-particle scattering S_X because it involves higher order clusters as scattering partners. More specifically, these can be interpreted as microscopic processes where a correlated two-particle quantity scatters from an incoherent carrier occupation $n^{\lambda} = f^{\mathrm{e}}$ or $n^{\lambda} = 1 - f^{\mathrm{h}}$. In our numerical calculations, we always include *the full structure* of $[D_X]_{\mathrm{inc}}$. This way, the analysis fully incorporates, e.g., the microscopic restrictions for exciton populations as a consequence of Pauli-blocking effects and scattering with electrons and holes.

When the carrier densities are relatively low, the dominant scattering contributions originate from those terms which contain a phase-space filling term $(1 - f^{\mathrm{e}} - f^{\mathrm{h}})$. This allows us to introduce the so-called *main-sum approximation* [17, 21], where only these dominant contributions of $[D_X]_{\mathrm{inc}}$ are included. This approach proves to be very useful once we look for analytic solutions to Eq. (10.287). For this reason, we explicitly present only the main-sum structure in the second and third lines of Eq. (10.287). These main-sum terms describe the attractive interaction between electrons and holes, allowing them to become truly bound electron–hole pairs, i.e., *incoherent excitons*.

The fourth line of Eq. (10.287) contains $G_{X,\mathrm{Coul}}$ and $G_{X,\mathrm{phon}}$ which are responsible for the generation of incoherent excitons from excitonic polarization. The remaining two-particle contributions are denoted as D_{rest} and contain the terms beyond the main-sum contributions. As a last contribution, the c_X dynamics contains T_X which symbolizes the three-particle Coulomb and phonon terms. As presented here, the c_X dynamics (10.287) is formally exact, and the accuracy of the numerical solutions depends only on the accuracy with which the three-particle correlation terms can be included to the analysis.

The dynamical equations for the correlations $c^{\mathbf{q},\mathbf{k}',\mathbf{k}}_{\lambda,\nu,\nu',\lambda'}$ are structurally similar to Eq. (10.287). In the numerical solutions, we treat all of these equations together with the corresponding equations for the singlets. This way, we fully include one- and two-particle correlations and obtain a closed set of equations providing a consistent description of optical excitations in semiconductors up to the level of three-particle correlations. In the following, we will introduce different levels of approximations for these triplet contributions. Our most sophisticated and still numerically feasible approximation describes the triplet terms at the level where we include microscopic scattering among singlets and doublets. As we will show, this is a very reasonable approximation for many interesting semiconductor excitation conditions.

10.7.2 Terahertz Spectroscopy of Excitons

In order to keep the analysis as simple as possible and to concentrate on the precise identification of genuine exciton populations, we focus here on a situation where all interband coherences have decayed, i.e., $\mathbf{P}$ and all other coherent quantities vanish. Furthermore, we derive the THz intraband dynamics from the original $(\mathbf{A} \cdot \mathbf{p})$-picture, Eq. (10.156), and consider only classical THz fields described by the vector potential $\langle \mathbf{A} \rangle \equiv \langle A(t) \rangle \mathbf{e}_A$. Under these conditions, the response of a semiconductor to a THz field follows from the current

$$J = \frac{1}{\mathsf{S}} \sum_{\mathbf{k}_\parallel, \lambda} \left[j_\lambda(\mathbf{k}_\parallel) - e^2 \langle A(t) \rangle / m_0 \right] f^\lambda_{\mathbf{k}_\parallel}, \tag{10.290}$$

with the free-electron mass m_0 and the current-matrix element

$$j_\lambda(\mathbf{k}_\parallel) \equiv -|e| \hbar \mathbf{k}_\parallel \cdot \mathbf{e}_A / m_\lambda, \tag{10.291}$$

where $\mathbf{e}_A$ is the polarization direction of the THz field. If we assume that the classical THz field propagates perpendicular to the QW or QWI system, the interaction of the carriers with the THz field is governed by the Hamiltonian

$$H_{\mathrm{THz}} = - \sum_{\mathbf{k}_\parallel} j_\lambda(\mathbf{k}_\parallel) a^\dagger_{\lambda, \mathbf{k}_\parallel} \, a_{\lambda, \mathbf{k}_\parallel} \langle A(t) \rangle, \tag{10.292}$$

as discussed in Kira et al. [17], Koch et al. [20], and Kira et al. [98]. It can be shown that the pure THz absorption properties follow entirely from the carrier-density-dependent part of J [98, 99], i.e.,

$$J_{\mathrm{THz}} = \frac{1}{\mathsf{S}} \sum_{\mathbf{k}, \lambda} j_\lambda(\mathbf{k}) f^\lambda_{\mathbf{k}}. \tag{10.293}$$

To compute J_{THz}, we have to evaluate the dynamics of the densities:

$$\frac{\partial}{\partial t} f^{\mathrm{e}}_{\mathbf{k}_\parallel} = -\frac{2}{\hbar} \mathrm{Im} \left[\sum_{\mathbf{q}_\parallel, \mathbf{k}'_\parallel} V_{\mathbf{k}'_\parallel + \mathbf{q}_\parallel - \mathbf{k}_\parallel} c_X^{\mathbf{q}_\parallel, \mathbf{k}'_\parallel, \mathbf{k}_\parallel} - \sum_{\mathbf{q}_\parallel, \mathbf{k}'_\parallel} V_{\mathbf{q}_\parallel} c^{\mathbf{q}_\parallel, \mathbf{k}'_\parallel, \mathbf{k}_\parallel}_{c,c,c,c} \right], \tag{10.294}$$

$$\frac{\partial}{\partial t} f^{\mathrm{h}}_{\mathbf{k}_\parallel} = +\frac{2}{\hbar} \mathrm{Im} \left[\sum_{\mathbf{q}_\parallel, \mathbf{k}'_\parallel} V_{\mathbf{k}'_\parallel + \mathbf{q}_\parallel - \mathbf{k}_\parallel} c_X^{-\mathbf{q}_\parallel, \mathbf{k}_\parallel, \mathbf{k}'_\parallel} - \sum_{\mathbf{q}_\parallel, \mathbf{k}'_\parallel} V_{\mathbf{q}_\parallel} c^{\mathbf{q}_\parallel, \mathbf{k}'_\parallel, \mathbf{k}_\parallel}_{v,v,v,v} \right], \tag{10.295}$$

where we have assumed that the incoherent and homogeneous carrier system interacts with a THz field while phonon-coupling effects are neglected for

simplicity. Equations (10.294) and (10.295) show that the single-particle densities do not couple directly to the THz light. Thus, THz absorption must involve at least two-particle correlations, which identifies *THz absorption as a uniquely qualified method to directly detect many-body correlations for incoherent quasi-particle excitations.*

Starting from Eq. (10.292), we can easily convince ourselves that also $c_{c,c,c,c}$ and $c_{v,v,v,v}$ are not directly coupled to the THz fields. Furthermore, we can show that the exciton correlation is directly driven by

$$\mathrm{i}\hbar \frac{\partial}{\partial t} c_{\mathrm{X}}^{\mathbf{q}_{\|},\mathbf{k}'_{\|},\mathbf{k}_{\|}}\bigg|_{\mathrm{THz}} = -j(\mathbf{k}'_{\|} + \mathbf{q}_{\|} - \mathbf{k}_{\|})\langle A(t)\rangle c_{\mathrm{X}}^{\mathbf{q}_{\|},\mathbf{k}'_{\|},\mathbf{k}_{\|}}, \tag{10.296}$$

where we have identified the reduced current-matrix element,

$$j(\mathbf{k}_{\|}) \equiv j_{\mathrm{e}}(\mathbf{k}_{\|}) + j_{\mathrm{h}}(\mathbf{k}_{\|}). \tag{10.297}$$

The THz contribution (10.296) now has to be added to the dynamics of c_X which satisfies an equation structurally similar to Eq. (10.287) and is discussed in detail in Kira and Koch [1].

In addition to the THz response from Eq. (10.296), the usual exciton dynamics and the build-up of correlations must be solved numerically from Eq. (10.287) when exciton formation is to be studied. In order to gain some analytical insights, we use a generalized exciton operator derived from a generalized Wannier equation in analogy to Section 10.5.2. The annihilation of an exciton in state λ and with center-of-mass momentum $\hbar\mathbf{q}_{\|}$ is then given by

$$X_{\lambda,\mathbf{q}_{\|}} \equiv \sum_{\mathbf{k}_{\|}} \varphi_{\lambda}^{\mathrm{R}}(\mathbf{k}_{\|}) a^{\dagger}_{v,\mathbf{k}_{\|}-\mathbf{q}_{\mathrm{h}}} a_{c,\mathbf{k}_{\|}+\mathbf{q}_{\mathrm{e}}}, \tag{10.298}$$

where we have introduced the abbreviations

$$\mathbf{q}_{\mathrm{e}} = \frac{m_{\mathrm{e}}}{m_{\mathrm{e}} + m_{\mathrm{h}}}\mathbf{q}_{\|}, \quad \mathbf{q}_{\mathrm{h}} = \frac{m_{\mathrm{h}}}{m_{\mathrm{e}} + m_{\mathrm{h}}}\mathbf{q}_{\|}. \tag{10.299}$$

The inverse transformation from the exciton to the electron–hole picture follows from

$$a^{\dagger}_{v,\mathbf{k}_{\|}-\mathbf{q}_{\mathrm{h}}} a_{c,\mathbf{k}_{\|}+\mathbf{q}_{\mathrm{e}}} = \sum_{\lambda} \varphi_{\lambda}^{\mathrm{L}}(\mathbf{k}_{\|}) X_{\lambda,\mathbf{q}_{\|}}. \tag{10.300}$$

Again, we use real-valued exciton functions in momentum space. Consequently, we do not have to keep track of complex conjugation.

The excitonic correlations are transformed into the exciton basis via

$$\Delta\left\langle X^{\dagger}_{\lambda,\mathbf{q}_{\|}} X_{\nu,\mathbf{q}_{\|}}\right\rangle = \sum_{\mathbf{k}_{\|},\mathbf{k}'_{\|}} \varphi_{\lambda}^{\mathrm{L}}(\mathbf{k}_{\|})\varphi_{\nu}^{\mathrm{L}}(\mathbf{k}'_{\|}) c_{\mathrm{X}}^{\mathbf{q}_{\|},\mathbf{k}'_{\|}-\mathbf{q}_{\mathrm{h}},\mathbf{k}_{\|}+\mathbf{q}_{\mathrm{e}}} \equiv \Delta N_{\lambda,\nu}(\mathbf{q}_{\|}), \tag{10.301}$$

$$c_{\mathrm{X}}^{\mathbf{q}_{\|},\mathbf{k}'_{\|}-\mathbf{q}_{\mathrm{h}},\mathbf{k}_{\|}+\mathbf{q}_{\mathrm{e}}} = \sum_{\lambda,\nu} \varphi_{\lambda}^{\mathrm{R}}(\mathbf{k}_{\|})\varphi_{\nu}^{\mathrm{R}}(\mathbf{k}'_{\|})\Delta N_{\lambda,\nu}(\mathbf{q}_{\|}), \tag{10.302}$$

such that their response to THz radiation is given by

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\Delta N_{\lambda,\nu}(\mathbf{q}_{\|})|_{\mathrm{THz}} = \sum_{\beta}\Big[J_{\lambda,\beta}\Delta N_{\beta,\nu}(\mathbf{q}_{\|}) - J_{\nu,\beta}\Delta N_{\lambda,\beta}(\mathbf{q}_{\|})\Big]\langle A(t)\rangle, \tag{10.303}$$

where we identified the transition-matrix element between two exciton states,

$$J_{\alpha,\beta} \equiv \sum_{\mathbf{k}_{\|}} \varphi_{\alpha}^{\mathrm{L}}(\mathbf{k}_{\|})j(\mathbf{k}_{\|})\varphi_{\beta}^{\mathrm{R}}(\mathbf{k}_{\|}). \tag{10.304}$$

The full correlation dynamics is obtained as Eq. (10.303) is added to Eq. (10.287). For our analytical evaluation, it is convenient to transform also Eq. (10.287) into the exciton basis which results in

$$\begin{aligned}\mathrm{i}\hbar\frac{\partial}{\partial t}\Delta N_{\lambda,\nu}(\mathbf{q}_{\|}) = {} & (E_{\nu} - E_{\lambda})\Delta N_{\lambda,\nu}(\mathbf{q}_{\|}) \\ & + (E_{\nu} - E_{\lambda})N_{\lambda,\nu}(\mathbf{q}_{\|})_{\mathrm{S}} + S_{\mathrm{coh}}^{\lambda,\nu}(\mathbf{q}_{\|}) \\ & + \mathrm{i}G^{\lambda,\nu}(\mathbf{q}_{\|}) + D_{\mathrm{rest}}^{\lambda,\nu}(\mathbf{q}_{\|}) + T^{\lambda,\nu}(\mathbf{q}_{\|}),\end{aligned} \tag{10.305}$$

where the incoherent part of the singlet scattering, S_X, in Eq. (10.287) produces a source

$$N_{\lambda,\nu}(\mathbf{q}_{\|})_{\mathrm{S}} \equiv \left\langle X_{\lambda,\mathbf{q}_{\|}}^{\dagger} X_{\nu,\mathbf{q}_{\|}} \right\rangle_{\mathrm{S}} = \sum_{\mathbf{k}_{\|}} \varphi_{\lambda}^{\mathrm{L}}(\mathbf{k}_{\|}) f_{\mathbf{k}_{\|}+\mathbf{q}_{\mathrm{e}}}^{\mathrm{e}} f_{\mathbf{k}_{\|}-\mathbf{q}_{\mathrm{h}}}^{\mathrm{h}} \varphi_{\nu}^{\mathrm{L}}(\mathbf{k}_{\|}). \tag{10.306}$$

This contribution has a finite value in the incoherent regime whenever we have any quasi-particle excitation in the system. Particularly, it drives exclusively the non-diagonal $\Delta\langle X_{\lambda}^{\dagger}X_{\nu}\rangle$ since it exists in Eq. (10.305) only when $\lambda \neq \nu$. In addition, phonon-assisted exciton formation can of course lead to diagonal excitonic populations:

$$\Delta N_{\lambda}(\mathbf{q}_{\|}) = \Delta N_{\lambda,\lambda}(\mathbf{q}_{\|}), \tag{10.307}$$

as was studied in detail in Hoyer et al. [21].

10.7.3 Linear Terahertz Response

We evaluate the excitonic signatures in the THz current by taking a time derivative of J_{THz}. Using Eqs. (10.293), (10.294), (10.295), and (10.302), we obtain

$$\frac{\partial}{\partial t}J_{\mathrm{THz}} = \frac{2}{\hbar}\mathrm{Im}\left[\frac{1}{\mathsf{S}}\sum_{\mathbf{q}_{\|},\mathbf{k}'_{\|},\mathbf{k}_{\|}} V_{\mathbf{q}_{\|}}\left(j_{\mathrm{e}}(\mathbf{k}_{\|})c_{c,c,c,c}^{\mathbf{q}_{\|},\mathbf{k}'_{\|},\mathbf{k}_{\|}} - j_{\mathrm{h}}(\mathbf{k}_{\|})c_{v,v,v,v}^{\mathbf{q}_{\|},\mathbf{k}'_{\|},\mathbf{k}_{\|}}\right)\right]$$
$$+\frac{1}{\hbar}\mathrm{Im}\left[\frac{1}{\mathsf{S}}\sum_{\mathbf{q}_{\|},\lambda,\nu}(E_\nu - E_\lambda)J_{\lambda,\nu}\Delta N_{\lambda,\nu}(\mathbf{q}_{\|})\right], \tag{10.308}$$

where the property (10.249) of the exciton states has been used to simplify the matrix elements related to the exciton contributions. In general, $c_{c,c,c,c}$ and $c_{v,v,v,v}$ provide electron and hole scattering to the THz currents, which essentially leads to a damping of J_{THz}.

At this stage, we can perform a full numerical analysis of Eqs. (10.303), (10.304), (10.305), (10.306), (10.307), and (10.308). Even though we do this and present the results later, we first want to gain some analytic insight into the THz response. For this purpose, *and not for the full numerical evaluations*, we now introduce a few simplifications that do not compromise the essential aspects of THz physics. First, we assume that the incoherent semiconductor state is quasi-stationary. This means that f^{e}, f^{h}, and c_X are known and stationary before the weak THz excitation of the system. Since such weak THz fields induce only small currents which are damped as a consequence of carrier scattering, it is reasonable to approximate the full microscopic scattering by a phenomenological damping. In other words, for the analytic evaluations we replace the contributions of $c_{c,c,c,c}$ and $c_{v,v,v,v}$ by $-\gamma_J J_{\mathrm{THz}}$ in Eq. (10.308). We also limit the investigations to the linear response. Here, the exciton correlation can be split into two parts,

$$\Delta N_{\lambda,\nu}(\mathbf{q}_{\|}) = \Delta N_{\lambda,\nu}(\mathbf{q}_{\|})_{(0)} + \Delta N_{\lambda,\nu}(\mathbf{q}_{\|})_{(1)}, \tag{10.309}$$

where $\Delta\langle N\rangle_{(0)}$ is the quasi-stationary exciton correlation and $\Delta\langle N\rangle_{(1)}$ is the linear response to $\langle A\rangle$.

Under these conditions, the exciton correlation dynamics can be linearized such that Eqs. (10.303) and (10.305) together produce

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\Delta N_{\lambda,\nu}(\mathbf{q}_{\|})_{(1)} = (E_\nu - E_\lambda - \mathrm{i}\gamma)\Delta N_{\lambda,\nu}(\mathbf{q}_{\|})_{(1)}$$
$$+\sum_{\beta}\left[J_{\lambda,\beta}\Delta N_{\beta,\nu}(\mathbf{q}_{\|})_{(0)} - J_{\nu,\beta}\Delta N_{\lambda,\beta}(\mathbf{q}_{\|})_{(0)}\right]\langle A(t)\rangle, \tag{10.310}$$

where the main-sum approximation has been used. Furthermore, in the THz-generated contributions, $\Delta\langle N\rangle_{(1)}$, we have replaced the influence of three-particle scattering by a constant dephasing rate [100].

Defining the total density of exciton correlations as

$$\Delta n_{\lambda,\nu}^{(j)} \equiv \frac{1}{\mathsf{S}}\sum_{\mathbf{q}_{|}}\Delta N_{\lambda,\nu}(\mathbf{q}_{\|})_{(j)} \tag{10.311}$$

we may sum Eq. (10.310) over $\mathbf{q}_{\|}$ and take the Fourier transformation to obtain

$$\hbar\omega\Delta n^{(1)}_{\lambda,\nu}(\omega) = (E_\nu - E_\lambda - \mathrm{i}\gamma)\Delta n^{(1)}_{\lambda,\nu}(\omega) + \sum_\beta \left[J_{\lambda,\beta}\Delta n^{(0)}_{\beta,\nu} - J_{\nu,\beta}\Delta n^{(0)}_{\lambda,\beta}\right]\langle A(\omega)\rangle \tag{10.312}$$

Note that $\Delta n^{(0)}_{\beta,\nu}$ is quasi-stationary such that only the Fourier transform of the THz field appears in the last term. In the same way, we Fourier transform also Eq. (10.308) to obtain

$$-\mathrm{i}\hbar\omega J_{\mathrm{THz}}(\omega) = -\gamma_J J_{\mathrm{THz}}(\omega) + \frac{1}{2\mathrm{i}}\sum_{\lambda,\nu}(E_\nu - E_\lambda)J_{\lambda,\nu}\left[\Delta n^{(1)}_{\lambda,\nu}(\omega) - \left(\Delta n^{(1)}_{\lambda,\nu}(-\omega)\right)^*\right], \tag{10.313}$$

where we replaced the microscopic scattering of the current by a decay constant γ_J and noticed that the quasi-stationary $\Delta n^{(0)}_{\lambda,\nu}$ cannot contribute to the current.

Equations (10.312) and (10.313) are now closed and yield the solution

$$J_{\mathrm{THz}}(\omega) = \frac{1}{\hbar\omega + \mathrm{i}\gamma_J} \times \sum_{\nu,\lambda}\left(S^{\nu,\lambda}(\omega)\Delta n^{(0)}_{\nu,\lambda} - \left[S^{\nu,\lambda}(-\omega)\Delta n^{(0)}_{\nu,\lambda}\right]^*\right)\langle A(\omega)\rangle \tag{10.314}$$

From this expression, we see that the THz current only depends on the initial state of the incoherent quasi-particle excitations, the spectrum of the THz field, and the generic THz response function:

$$S^{\nu,\lambda}(\omega) = \sum_\beta \frac{(E_\beta - E_\nu)J_{\nu,\beta}J_{\beta,\lambda}}{E_\beta - E_\nu - \hbar\omega - \mathrm{i}\gamma} \tag{10.315}$$

The denominator of this response function introduces resonances corresponding to transitions between different exciton states, whereas the product of the matrix elements $J_{\nu,\beta}J_{\beta,\lambda}$ provides the selection rules.

Just as in the case of linear interband absorption, the result (10.314) can be directly applied to produce the linear susceptibility:

$$\chi_{\mathrm{THz}} \equiv \frac{P_{\mathrm{THz}}(\omega)}{\varepsilon_0\langle E(\omega)\rangle} = \frac{J_{\mathrm{THz}}(\omega)}{\varepsilon_0\omega^2\langle A(\omega)\rangle}, \tag{10.316}$$

where we used the general relations, $\langle E(t)\rangle = -\frac{\partial}{\partial t}\langle A(t)\rangle$ and $J_{\mathrm{THz}}(t) \equiv \frac{\partial}{\partial t}P_{\mathrm{THz}}(t)$ to evaluate $\langle E(\omega)\rangle = \mathrm{i}\omega\langle A(\omega)\rangle$ and $P_{\mathrm{THz}}(\omega) = \frac{1}{\omega}J_{\mathrm{THz}}(\omega)$, respectively. Since the QW is thin compared with the THz wavelength (we have assumed that the

planar confinement is much smaller than the optical wavelength), we may compute the THz absorption from the formula

$$\alpha_{\mathrm{THz}}(\omega) = \frac{\omega}{nc}\mathrm{Im}[\chi_{\mathrm{THz}}(\omega)], \tag{10.317}$$

which provides a good approximation for small $|\chi_{\mathrm{THz}}| \ll nc/\omega$. As we insert the result (10.314) into Eq. (10.317), we find

$$\alpha_{\mathrm{THz}}(\omega) = \mathrm{Im}\left[\sum_{\nu,\lambda} \frac{S^{\nu,\lambda}(\omega)\Delta n^{(0)}_{\nu,\lambda} - \left[S^{\nu,\lambda}(-\omega)\Delta n^{(0)}_{\nu,\lambda}\right]^*}{\varepsilon_0 nc\omega(\hbar\omega + \mathrm{i}\gamma_J)}\right] \tag{10.318}$$

which gives the THz absorption from a generic incoherent quasi-particle state.

To gain some more detailed insights, we first analyze the THz absorption for the limiting case where only diagonal correlations exist, i.e., $\Delta n^{(0)}_{\nu,\lambda} = \delta_{\nu,\lambda}\Delta n^{(0)}_{\nu,\nu} \equiv \delta_{\nu,\lambda}\Delta n_\nu$. In this situation, Eq. (10.318) reduces to

$$\alpha_{\mathrm{atom}}(\omega) = \frac{\omega}{\varepsilon_0 nc}\mathrm{Im}\left[\sum_\nu \left(S^\nu_{\mathrm{atom}}(\omega) - \left[S^\nu_{\mathrm{atom}}(-\omega)\right]^*\right)\Delta n_\nu\right], \tag{10.319}$$

$$S^\nu_{\mathrm{atom}}(\omega) = \sum_\beta \frac{|D_{\nu,\beta}|^2}{E_\beta - E_\nu - \hbar\omega - \mathrm{i}\gamma}. \tag{10.320}$$

Here, we defined the excitonic dipole matrix element

$$D_{\lambda,\nu} \equiv \langle\varphi^{\mathrm{L}}_\lambda|e\mathbf{r}\cdot\mathbf{e}_{\mathrm{P}}|\varphi^{\mathrm{R}}_\nu\rangle = \frac{\mathrm{i}\hbar}{E_\nu - E_\lambda}J_{\lambda,\nu}, \tag{10.321}$$

using the general connection of dipole- and current-matrix elements [4]. We have also assumed that the $(E_\nu - E_\lambda)$ in Eq. (10.321) as well as in the numerator of Eq. (10.315) can be replaced by $\hbar\omega$ due to the narrow enough Lorentzian resonances in $S^{\lambda,\nu}$. With these assumptions, which are typical in atom optics, we find that our THz analysis produces an atom-like absorption spectrum for the case where different atomic levels are populated according to Δn_ν. This result clearly establishes the close relation between excitonic THz and atomic spectroscopy helping us to give physical support to our concept of exciton populations.

Based on the results discussed so far, we may anticipate that electrons and holes must first come close to each other in real space, before they can form bound excitons. Thus, it is natural to follow how the electron–hole-pair correlation function evolves in time as the exciton formation proceeds. Figure 10.10 shows a computed sequence of $\Delta g_{\mathrm{eh}}(r)$ as a function of electron–hole distance r for a low carrier density of $n^{\mathrm{e/h}} = 2\times 10^4\,\mathrm{cm}^{-1}$. Already at early times around $t = 0.5$ ps, we see that the probability of finding electrons and holes close to each other increases as a consequence of the Coulomb attraction. We also

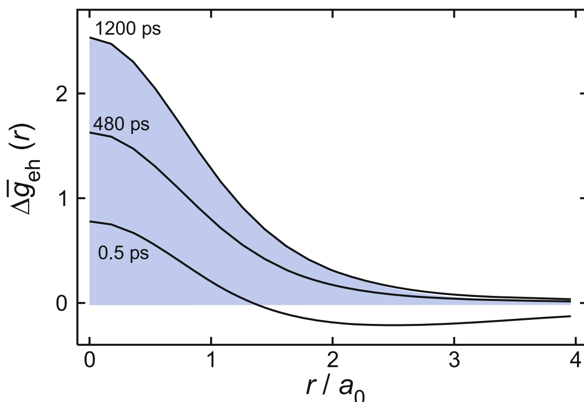


Fig. 10.10 Pair correlation function $\Delta g_{eh}(r)$ for the lattice temperature of $T = 10$ K and carrier density $n = 2 \times 10^4 \text{cm}^{-1}$ at different times. The absolute square of the 1*s*-exciton wavefunction is shown as *shaded area*. (From Hoyer et al. [21])

notice that the correlated Δg_{eh} at early times has clearly negative parts indicating a transient depletion caused by the overall reduction of the electron–hole separation. This form corresponds to the generation of a correlated electron–hole plasma [21]. At later times, $\Delta g_{eh}(r)$ becomes entirely positive and grows linearly in magnitude. In particular, $\Delta g_{eh}(r)$ then assumes the shape of the probability distribution of 1s excitons (shaded area). Thus, the formation of truly bound 1s excitons proceeds in the sequence that (i) a correlated plasma is built up on a sub-picosecond time scale due to Coulomb interaction and (ii) phonon-assisted scattering forms excitons out of the correlated plasma on a nanosecond time scale.

To illustrate how the exciton formation can be directly detected experimentally, we compute the THz absorption spectrum resulting from non-resonant excitation with a 500 fs excitation pulse energetically 16 meV above the 1s-exciton resonance. The pulse intensity is chosen such that it generates a moderate 6×10^4 cm^{-1} carrier density. In Fig. 10.11, we see that the computed

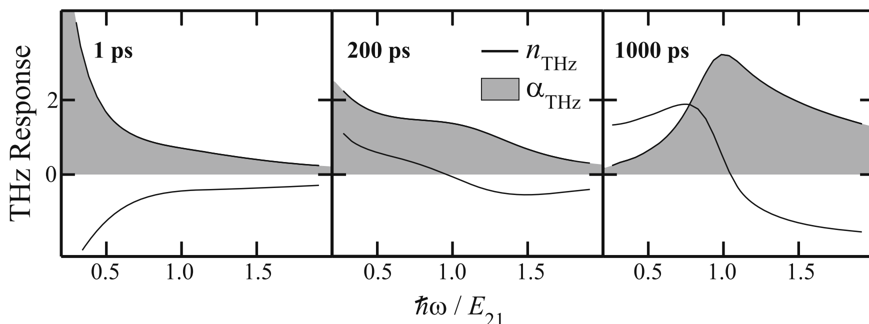


Fig. 10.11 Computed THz absorption (*shaded area*) and refractive index changes (*solid line*) for different THz probe delays after non-resonant excitation. Here, $E_{2p-1s} = 5$ is the energy difference between 1 s and 2 p states. (From Kira and Koch [97])

$\alpha_{\mathrm{THz}}(\omega)$ is very broad and shows no resonances at 1 ps after the excitation. Even after 200 ps, the THz response has changed only slightly due to the slow phonon scattering from electron–hole plasma to excitons. However, roughly 1 ns after the excitation, $\alpha_{\mathrm{THz}}(\omega)$ develops a pronounced resonance at the energy corresponding exactly to the difference between the two lowest exciton states. The asymmetric shape of $\alpha_{\mathrm{THz}}(\omega)$ is a consequence of transitions between the lowest and all other exciton states. These results are in good qualitative agreement with recent experiments [101].

References

1. M. Kira and S. W. Koch. Many-body correlations and excitonic effects in semiconductor spectroscopy. *Prog. Quantum Electron.* **30**:155–296, 2006.
2. M. Born and R. Oppenheimer. Qantum theory of molecules. *Ann. Phys.*, **84**:457–484, 1927.
3. C. Cohen-Tannoudji, J. Dupont-Roc, and G. Grynberg. *Photons and Atoms*. Wiley, New York, 3. edition, 1989.
4. H. Haug and S. W. Koch. *Quantum Theory of the Optical and Electronic Properties of Semiconductors*. World Scientific Publ., Singapore, 4. edition, 2004.
5. E. Merzbacher. *Quantum Mechanics*. Wiley, New York, 1. edition, 1961.
6. L. I. Schiff. *Quantum Mechanics*. MacGraw-Hill, New York, 3. edition, 1968.
7. M. Kira, F. Jahnke, W. Hoyer, and S. W. Koch. Quantum theory of spontaneous emission and coherent effects in semiconductor microstructures. *Prog. Quantum Electron.*, **23**:189–279, 1999.
8. M. Goeppert Mayer. Elementary processes with two-quantum transitions. *Ann. d. Physik*, **9**:273, 1931.
9. J. Cízek. On correlation problem in atomic and molecular systems. Calculation of wavefunction components in ursell-type expansion using quantum-field theoretical methods. *J. Chem. Phys.*, **45**:4256, 1966.
10. J. Fricke. Transport equations including many-particle correlations for an arbitrary quantum system: A general formalism. *Ann. Phys.*, **252**(2):479–498, 1996.
11. F. E. Harris, H. J. Monkhorst, and D. L. Freeman. *Algebraic and Diagrammatic Methods in Many-Fermion Theory*. Oxford Press, New York, 1. edition, 1992.
12. W. Hoyer, M. Kira, and S. W. Koch. Cluster expansion in semiconductor quantum optics. In K. Morawetz, editor, *Nonequilibrium Physics at Short Time Scales*, pages 309–335. Springer Verlag, Berlin, 2004.
13. M. Kira and S. W. Koch. Microscopic theory of optical excitations, photoluminescence, and terahertz response in semiconductors. *Eur. J. Phys. D*, **36**:143–157, 2005.
14. M. Kira, W. Hoyer, and S. W. Koch. Excitons and luminescence in semiconductor heterostructures. *Nonlinear Opt.*, **29**:481–489, 2002.
15. G. D. Purvis and R. J. Bartlett. A full coupled-cluster singles and doubles model: The inclusion of disconnected triples. *J. Chem. Phys.*, **76**:1910–1918, 1982.
16. H. W. Wyld and B. D. Fried. Quantum mechanical kinetic equations. *Ann. phys.*, **23**:374–389,1963.
17. M. Kira, W. Hoyer, T. Stroucken, and S. W. Koch. Exciton formation in semiconductors and the influence of a photonic environment. *Phys. Rev. Lett.*, **87**:176401, 2001.
18. M. Kira, W. Hoyer, S. W. Koch, P. Brick, C. Ell, M. Hübner, G. Khitrova, and H. M. Gibbs. Quantum correlations in semiconductor microcavities. *Semicond. Sci. Technol.*, **18**:S405–S410, 2003.
19. S. W. Koch and M. Kira. Excitons in semiconductors. In H. Kalt and M. Hetterich, editors, *Optics of Semiconductors and Their Nanostructures – Springer Series in Solid-State Sciences Vol. 146*, pages 1–18. Springer Verlag, Berlin, 2004.

20. S. W. Koch, M. Kira, G. Khitrova, and H. M. Gibbs. Excitons in new light. *Nat. Mater.*, **5**:523–531, 2006.
21. W. Hoyer, M. Kira, and S. W. Koch. Influence of Coulomb and phonon interaction on the exciton formation dynamics in semiconductor heterostructures. *Phys. Rev. B*, **67**:155113, 2003.
22. S. Siggelkow, W. Hoyer, M. Kira, and S. W. Koch. Exciton formation and stability in semiconductor heterostructures. *Phys. Rev. B*, **69**:073104, 2004.
23. V. S. Filinov, W. Hoyer, M. Bonitz, M. Kira, V. E. Fortov, and S. W. Koch. Spontaneous emission of semiconductors in the Wigner approach. *J. Opt. B*, **5**:S299–S305, 2003.
24. S. W. Koch, M. Kira, W. Hoyer, and V. S. Filinov. Exciton ionization in semiconductors. *Phys. Stat. Sol. B*, **238**:404–410, 2003.
25. M. Hentschel, R. Kienberger, C. Spielmann, G. A. Reider, N. Milosevic, T. Brabec, P. Corkum, U. Heinzmann, M. Drescher, and F. Krausz. Attosecond metrology. *Nature*, **414**:509–513, 2001.
26. D. J. Jones, S. A. Diddams, J. K. Ranka, A. Stenz, R. S. Windeler, J. L. Hall, and S. T. Cundiff. Carrier-envelope phase control of femtosecond mode-locked lasers and direct optical frequency synthesis. *Science*, **288**:635–639,2000.
27. J. Shah. *Ultrafast Spectroscopy of Semiconductors and Semiconductor Nanostructures – Springer Series in Solid State Sciences, Vol. 115*. Springer Verlag, New York, 2. edition, 1999.
28. T. Udem, R. Holzwarth, and T. W. Hänsch. Optical frequency metrology. *Nature*, **416**:233–237, 2002.
29. G. W. Fehrenbach, W. Schäfer, J. Treusch, and R. G. Ulbrich. Transient optical spectra of a dense exciton gas in a direct-gap semiconductor. *Phys. Rev. Lett.*, **57**:1281–1284, 1982.
30. H. M. Gibbs, A. C. Gossard, S. L. McCall, A Passner, W. Wiegmann, and T. N. C. Venkatesan. Saturation of the free exciton resonance in GaAs. *Solid State Commun.*, **30**:271–275, 1979.
31. Y. H. Lee, A. Chavezpirson, S. W. Koch, H. M. Gibbs, S. H. Park, J. Morhange, A. Jeffery, N. Peyghambarian, L. Banyai, A. C. Gossard, and W. Wiegmann. Room-temperature optical nonlinearities in GaAs. *Phys. Rev. Lett.*, **57**:2446–2449, 1986.
32. V. G. Lysenko and V. I. Revenko. Exciton spectrum in case of high-density non-equilibrium carriers in CdS crystals. *Fizika Tverdogo Tela*, **20**:2144–2147, 1978.
33. T. B. Norris, J.-K. Rhee, C.-Y. Sung, Y. Arakawa, M. Nishioka, and C. Weisbuch. Time-resolved vacuum rabi oscillations in a semicondcutor quantum microcavity. *Phys. Rev. B*, **50**:14663–14666, 1994.
34. S. Schmitt-Rink, D. S. Chemla, and D. A. B. Miller. Linear and nonlinear optical properties of semiconductor quantum wells. *Adv. Phys.*, **38**:89–188, 1989.
35. C. Weisbuch, M. Nishioka, A. Ishikawa, and Y. Arakawa. Observation of the coupled exciton-photon mode splitting in a semiconductor quantum microcavity. *Phys. Rev. Lett.*, **69**:3314–3317, 1992.
36. D. S. Chemla and J. Shah. Many-body and correlation effects in semiconductors. *Nature*, **411**:549–557, 2001.
37. S. T. Cundiff, M. Koch, W. H. Knox, J. Shah, and W. Stolz. Optical coherence in semiconductors: Strong emission mediated by nondegenerate interactions. *Phys. Rev. Lett.*, **77**:1107–1110, 1996.
38. P. Kner, W. Schäfer, R. Lövenich, and D. S. Chemla. Coherence of four-particle correlations in semiconductors. *Phys. Rev. Lett.*, **81**:5386–5389, 1998.
39. J. Kuhl. Optical dephasing of excitons in iii–v semiconductors. In R. T. Phillips, editor, *Coherent Optical Interactions in Semiconductors*, pages 1–31. Plenum Press, New York, 1994.
40. T. Rappen, G. Mohs, and M. Wegener. Polariton dynamics in quantum wells studied by femtosecond four-wave mixing. *Phys. Rev. B*, **47**:9658–9662, 1993.

41. W. Schäfer, D. S. Kim, J. Shah, T. C. Damen, J. E. Cunningham, L. N. Pfeiffer, K. W. Goossen, and K. Köhler. Femtosecond coherent fields induced by many-particle correlations in transient four-wave-mixing. *Phys. Rev. B*, **53**:16429–16443, 1996.
42. W. Schäfer, R. Lövenich, N. A. Fromer, and D. S. Chemla. From coherently excited highly correlated states to incoherent relaxation processes in semiconductors. *Phys. Rev. Lett.*, **86**:344–347, 2001.
43. L. Schultheis, M. D. Sturge, and J. Hegarty. Photon-echoes from two-dimensional excitons in GaAs–AlGaAs quantum wells. *Appl. Phys. Lett.*, **47**:995–997, 1985.
44. L. Schultheis, J. Kuhl, A. Honold, and C. W. Tu. Ultrafast phase relaxation of excitons via exciton–exciton and exciton–electron collisions. *Phys. Rev. Lett.*, **57**:1635–1638, 1986.
45. A. L. Smirl. The vectorial dynamics of coherent emission from excitons. In K.-T. Tsen, editor, *Ultrafast Phenomena in Semiconductors*, pages 443–507. Springer Verlag, New York, 2001.
46. H. Stolz. *Time Resolved Light Scattering from Excitons*. Springer Verlag, Berlin, 1994.
47. M. Lindberg and S. W. Koch. Effective Bloch equations for semiconductors. *Phys. Rev. B*, **38**:3342–3350, 1988.
48. H. Haug and S. Schmitt-Rink. Electron theory of the optical properties of laser excited semiconductors. *Prog. Quantum Electron.*, **9**:3–100, 1984.
49. L. V. Keldysh and Y. V. Kopaev. Possible instability of the semimetal state toward coulomb interaction. *Sov. Phys. Solid State*, **6**:2219–2224, 1965.
50. C. Klingshirn and H. Haug. Optical properties of highly excited direct gap semiconductors. *Phys. Rep.*, **70**:315–410, 1981.
51. R. J. Elliott. Theory of excitons. In C. G. Kuper and G. D. Whitefield, editors, *Polarons and Excitons*, pages 269–293. Oliver and Boyd, Edinburgh, 1963.
52. F. Jahnke, M. Kira, and S. W. Koch. Linear and nonlinear optical properties of quantum confined excitons in semiconductor microcavities. *Z. Physik B*, **104**:559–572, 1997.
53. S. W. Koch, N. Peyghambarian, and M. Lindberg. Transient and steady-state optical nonlinearities in semiconductors. *J. Phys. C: Solid State Phys.*, **21**:5229–5249, 1988.
54. B. Mieck, H. Haug, W. A. Hügel, M. F. Heinrich, and M. Wegener. Quantum-kinetic dephasing in resonantly excited semiconductor quantum wells. *Phys. Rev. B*, **62**:2686–2695, 2000.
55. T. Rappen, U. G. Peter, M. Wegener, and W. Schäfer. Polarization dependence of dephasing processes – A probe for many-body effects. *Phys. Rev. B*, **49**:10774–10777, 1994.
56. S. Chatterjee, C. Ell, S. Mosor, G. Khitrova, H. M. Gibbs, W. Hoyer, M. Kira, S. W. Koch, J. P. Prineas, and H. Stolz. Excitonic photoluminescence in semiconductor quantum wells: Plasma versus excitons. *Phys. Rev. Lett.*, **92**:067402, 2004.
57. W. W. Chow and S. W. Koch. *Semiconductor Laser Fundamentals*. Springer Verlag, New York, 1. edition, 1999.
58. F. Jahnke, M. Kira, S. W. Koch, G. Khitrova, E. K. Lindmark, T. R. Nelson Jr., D. V. Wick, J. D. Berger, O. Lyngnes, H. M. Gibbs, and K. Tai. Excitonic nonlinearities of semiconductor microcavities in the nonperturbative regime. *Phys. Rev. Lett.*, **77**:5257–5260, 1996.
59. W. Hoyer, M. Kira, and S. W. Koch. Influence of bound and unbound electron–hole-pair populations and interaction effects on the excitonic luminescence in semiconductor quantum wells. *Cond-Mat*, 0604349, 2006.
60. N. F. Mott. The transition to the metallic state. *Philos. Mag.*, **6**:287–309, 1961.
61. W. W. Chow, S. W. Koch, and M. Sargent III. *Semiconductor-Laser Physics*. Springer Verlag, Berlin, corrected second printing 1997 edition, 1994.
62. W. W. Chow, A. F. Wright, A. Girndt, F. Jahnke, and S. W. Koch. Microscopic theory of gain in an inhomogeneously broadened ingan/algan quantum-well laser. *Appl. Phys. Lett.*, **71**:2608–2610, 1997.
63. J. Hader, S. W. Koch, and J. V. Moloney. Microscopic theory of gain and spontaneous emission in GaInNAs laser material. *Solid State Electron.*, **47**:513–521, 2003.

64. J. Hader, J. V. Moloney, S. W. Koch, and W. W. Chow. Microscopic modelling of gain and luminescence in semiconductors. *J. Sel. Top. Quant. Electron.*, **9**:688–697, 2003.
65. C. Ell, P. Brick, M. Hübner, E. S. Lee, O. Lyngnes, J. P. Prineas, G. Khitrova, H. M. Gibbs, M. Kira, F. Jahnke, S. W. Koch, D. G. Deppe, and D. L. Huffaker. Quantum correlations in the nonperturbative regime of semiconductor microcavities. *Phys. Rev. Lett.*, **85**:5392–5395,2000.
66. W. Hoyer, M. Kira, S. W. Koch, H. Stolz, S. Mosor, J. Sweet, C. Ell, G. Khitrova, and H. M. Gibbs. Entanglement between a photon and a quantum well. *Phys. Rev. Lett.*, **93**:067401, 2004.
67. M. Kira, F. Jahnke, and S. W. Koch. Microscopic theory of excitonic signatures in semiconductor photoluminescence. *Phys. Rev. Lett.*, **81**:3263–3266, 1998.
68. M. Kira, F. Jahnke, and S. W. Koch. Quantum theory of secondary emission in optically excited semiconductor quantum wells. *Phys. Rev. Lett.*, **82**:3544–3547, 1999.
69. Y.-S. Lee, T. B. Norris, M. Kira, F. Jahnke, S. W. Koch, G. Khitrova, and H. M. Gibbs. Quantum correlations and intraband coherences in semiconductor cavity QED. *Phys. Rev. Lett.*, **83**:5338–5341, 1999.
70. S. Hoffmann, M. Hofmann, E. Bründermann, M. Havenith, M. Matus, J. V. Moloney, A. S. Moskalenko, M. Kira, S. W. Koch, S. Saito, and K. Sakai. Four-wave mixing and direct terahertz emission with two-color semiconductor lasers. *Appl. Phys. Lett.*, **84**:3585–3587, 2004.
71. S. Hoffmann, M. Hofmann, M. Kira, and S. W. Koch. Two-colour diode lasers for generation of THz radiation. *Semicond. Sci. Technol.*, **20**:205–210, 2005.
72. W. Hoyer, A. Knorr, J. V. Moloney, E. M. Wright, M. Kira, and S. W. Koch. Photoluminescence and terahertz emission from femtosecond laser-induced plasma channels. *Phys. Rev. Lett.*, **94**:115004, 2005.
73. M. Richter, M. Schaarschmidt, A. Knorr, W. Hoyer, J. V. Moloney, E. M. Wright, M. Kira, and S. W. Koch. Quantum theory of incoherent THz-emission of an interacting electron–ion plasma. *Phys. Rev. A*, **71**:053819, 2005.
74. W. Hoyer, C. Ell, M. Kira, S. W. Koch, S. Chatterjee, S. Mosor, G. Khitrova, H. M. Gibbs, and H. Stolz. Many-body dynamics and exciton formation studied by time-resolved photoluminescence. *Phys. Rev. B*, **72**:075324, 2005.
75. I. Galbraith, R. Chari, S. Pellegrini, P. J. Phillips, C. J. Dent, A. F. G. van der Meer, D. G. Clarke, A. K. Kar, G. S. Buller, C. R. Pidgeon, B. N. Murdin, J. Allam, and G. Strasser. Excitonic signatures in the photoluminescence and terahertz absorption of a GaAs/Al_xGa_{1-x}As multiple quantum well. *Phys. Rev. B*, **71**:073302, 2005.
76. M. Kira, F. Jahnke, S. W. Koch, J. D. Berger, D. V. Wick, T. R. Nelson Jr., G. Khitrova, and H. M. Gibbs. Quantum theory of nonlinear semiconductor microcavity luminescence explaining "Boser" experiments. *Phys. Rev. Lett.*, **79**:5170–5173, 1997.
77. G. Khitrova, H. M. Gibbs, F. Jahnke, M. Kira, and S. W. Koch. Nonlinear optics of normal-mode-coupling semiconductor microcavities. *Rev. Mod. Phys.*, **71**:1591–1639, 1999.
78. W. Chow, M. Kira, and S. W. Koch. Microscopic theory of optical nonlinearities and spontaneous emission lifetime in group iii nitride quantum wells. *Phys. Rev. B*, **60**:1947–1952, 1999.
79. K. Hantke, J. D. Heber, C. Schlichenmaier, A. Thränhardt, T. Meier, B. Kunert, K. Volz, W. Stolz, S. W. Koch, and W. W. Rühle. Time-resolved photoluminescence of type-i and type-ii (GaIn)As/Ga(NAs) heterostructures. *Phys. Rev. B*, **71**:165320, 2005.
80. R. F. Schnabel, R. Zimmermann, D. Bimberg, H. Nickel, R. Lösch, and W. Schlapp. Influence of exciton localization on recombination line shapes: In_xGa_{1-x}As/GaAs quantum wells as a model. *Phys. Rev. B*, **46**:9873–9876, 1992.
81. J. Szczytko, L. Kappei, J. Berney, F. Morier-Genoud, M. T. Portella-Oberli, and B. Deveaud. Determination of the exciton formation in quantum wells from time-resolved interband luminescence. *Phys. Rev. Lett.*, **93**:137401, 2004.

82. J. Szczytko, L. Kappei, J. Berney, F. Morier-Genoud, M. T. Portella-Oberli, and B. Deveaud. Origin of excitonic luminescence in quantum wells: Direct comparison of the exciton population and coulomb correlated plasma models. *Phys. Rev. B*, **71**:195313, 2005.
83. G. Chen, N. H. Bonadeo, D. G. Steel, D. Gammon, D. S. Katzer, D. Park, and L. J. Sham. Optically induced entanglement of excitons in a single quantum dot. *Science*, **289**:1906–1909,2000.
84. X. Q. Li, Y. W. Wu, D. G. Steel, D. Gammon, T. H. Stievater, D. S. Katzer, D. Park, C. Piermarocchi, and L. J. Sham. An all-optical quantum gate in a semiconductor quantum dot. *Science*, **301**:809–811, 2003.
85. P. Michler, A. Kiraz, C. Becher, W. V. Schoenfeld, P. M. Petroff, L. D. Zhang, E. Hu, and A. Imamoglu. A quantum dot single-photon turnstile device. *Science*, **290**:2282–2285,2000.
86. W. Hoyer, M. Kira, and S. W. Koch. Quantum-optical effects in semiconductors. *Festkörperprobleme (Adv. Solid State Phys.)*, **42**:55, 2002.
87. V. M. Agranovich and O. A. Dubowskii. Effect of retarded interaction of exciton spectrum in 1-dimensional and 2-dimenstional crystals. *JETP Lett.*, **3**:223, 1966.
88. F. Tassone, F. Bassani, and L. C. Andreani. Quantum-well reflectivity and exciton–polariton dispersion. *Phys. Rev. B*, **45**:6023–6030, 1992.
89. B. G. Englert, M. O. Scully, and H. Walther. Complementarity and uncertainty. *Nature*, **375**:367–368, 1995.
90. M. O. Scully, B. G. Englert, and H. Walther. Quantum optical tests of complementarity. *Nature*, **351**:111–116, 1991.
91. E. P. Storey, S. M. Tan, M. J. Collett, and D. F. Walls. Path detection and the uncertainty principle. *Nature*, **367**:626–628, 1994.
92. P. Bozsoki, P. Thomas, M. Kira, W. Hoyer, T. Meier, S.W. Koch, K. Maschke, I. Varga, and H. Stolz. Characterization of disorder in semiconductors via single–photon interferometry. *Phys. Rev. Lett.*, accepted, 2006.
93. J. Cerne, J. Kono, M. S. Sherwin, M. Sundaram, A. C. Gossard, and G. E. W. Bauer. Terahertz dynamics of excitons in GaAs/AlGaAs quantum wells. *Phys. Rev. Lett.*, **77**:1131–1134, 1996.
94. E. M. Gershenzon, G. N. Goltsman, and M. G. Ptitsina. Investigation of free excitons in Ge and their condensation at submillimeter waves. *Zhurnal Eksperimentalnoi I Teoreticheskoi Fiziki*, **70**:224–234, 1976.
95. R. M. Groeneveld and D. Grischkowsky. Picosecond time-resolved far-infrared experiments on carriers and excitons in GaAs–AlGaAs multiple-quantum wells. *J. Opt. Soc. Am. B*, **11**:2502–2507, 1994.
96. T. Timusk, R. Navarro, N. O. Lipari, and M. Altarelli. Far-infrared absorption by excitons in silicon. *Solid State Commun.*, **25**:217–219, 1978.
97. M. Kira and S. W. Koch. Exciton-population inversion and terahertz gain in resonantly excited semiconductors. *Phys. Rev. Lett.*, **93**:076402, 2004.
98. M. Kira, W. Hoyer, and S. W. Koch. Microscopic theory of the semiconductor terahertz response. *Phys. Stat. Sol. B*, **238**:443–450, 2003.
99. M. Kira, W. Hoyer, and S. W. Koch. Terahertz signatures of the exciton formation dynamics in non-resonantly excited semiconductors. *Solid State Commun.*, **129**:733–736, 2004.
100. M. Kira, W. Hoyer, S.W. Koch, Y.-S. Lee, T. B. Norris, G. Khitrova, and H. M. Gibbs. Incoherent pulse generation in semiconductor microcavities. *Phys. Stat. Sol. C*, **0**:1397–1400, 2003.
101. R. A. Kaindl, M. A. Carnahan, D. Hagele, R. Lovenich, and D. S. Chemla. Ultrafast terahertz probes of transient conducting and insulating phases in an electron-hole gas. *Nature*, **423**:734–738, 2003.

Chapter 11
Photonic Crystals: Physics, Fabrication, and Devices

Wei Jiang and Michelle L. Povinelli

Abstract We review basic physics of photonic crystals, discuss the relevant fabrication techniques, and summarize important device development in the past two decades. First, photonic band structures of photonic crystals and the origin of the photonic band gap are analyzed. Fundamental photonic crystal structures, such as surfaces, slabs, and engineered defects that include cavities and waveguides, are examined. Applications at visible and infrared wavelengths require photonic crystals to have submicron features, sometimes with precision down to the nanoscale. Common fabrication methods that have helped make such exquisite structures will be reviewed. Lastly, we give a concise account of key advances in photonic crystal-based lasers, light-emitting devices, modulators, optical filters, superprism-based demultiplexers and sensors, and negative index materials. Electron-beam nanolithography has enabled major research progress on photonic crystal devices in the last decade, leading to significant reduction of size and/or power dissipation in devices such as lasers and modulators. With deep ultraviolet (DUV) lithography, these devices may one day be manufactured with the prevalent CMOS technology at affordable cost.

11.1 Introduction

The concepts of electronic band structure and electronic band gaps revolutionized the scientific study of crystalline solids. This understanding gave birth to semiconductor and integrated electronic devices that have fundamentally changed our life and society. In the same way as the periodic lattice of a crystalline

W. Jiang
Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ 08854 and Omega Optics, Inc., Austin, Texas 78758, USA
e-mail: wjiangnj@ece.rutgers.edu

M.L. Povinelli
Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089
e-mail: povinell@usc.edu

A. Korkin, F. Rosei (eds.), *Nanoelectronics and Photonics*,
DOI: 10.1007/978-0-387-76499-3_11, © Springer Science+Business Media, LLC 2008

solid results in band gaps in the electronic energy spectrum, a periodic dielectric structure may give rise to gaps in the *photonic* frequency spectrum, or equivalently the energy spectrum of photons. Such gaps are called *photonic band gaps* (PBGs), and the structure is called a *photonic crystal*. Today, our growing understanding of photonic crystals is revolutionizing the design of optical devices, as exemplified in the work on lasers and modulators presented in this chapter.

The systematic understanding of photonic band structure has developed only in the last 20 years. In 1987, Eli Yablonovitch and Sajeev John independently recognized the significance of the photonic band gap while studying two apparently disparate topics, laser cavities [1] and localization in disordered dielectric media [2]. Yablonovitch, then at Bell Communications Research, considered dielectric structures with periodicity on the wavelength scale. He proposed that the broadband spontaneous emission of atoms in such a structure would be prohibited in a photonic band gap. As a result, spontaneous emission loss would be reduced, and a laser cavity constructed in the photonic crystal could achieve a vanishing threshold [1]. Meanwhile, John, then at Princeton, was studying the problem of light localization in a moderately disordered dielectric structure [2]. He realized that the photonic band gap of a nearly periodic dielectric structure could enhance the light localization, supplementing the well-known Anderson localization mechanism due to structural disorder. Thus, the study of photonic crystals was initiated technologically and scientifically.

The early development of photonic crystals focused on fabricating three-dimensional (3D) photonic crystals at microwave wavelengths [3,4]. Because the lattice constant of a photonic crystal is proportional to its operating wavelength the feature sizes of microwave photonic crystals were large enough to be amenable to machining. Significant challenges emerged, however, in scaling down 3D crystals to optical wavelengths. Most photonic crystal devices at visible or near-infrared (NIR) wavelengths are instead based on simpler, 2D periodic photonic crystals.

During the last two decades, photonic crystal research has expanded and flourished. This chapter is intended to briefly summarize key device research, augmented by certain scientific developments that enabled the device concepts. We will first give a concise introduction to the concept of bands and band gaps of photonic crystals, followed by optical properties of waveguides and microcavities. We then discuss photonic crystal surfaces and the formulation of a general transmission theory for photonic crystals. Fabrication methods will be introduced prior to moving into the sections on various devices. Subsequently, we present device research on photonic crystal lasers, filters, modulators, followed by discussion of "superprism"-based devices and negative index materials. Lastly, we summarize the advances of photonic crystal research and reflect on future directions. Due to the tutorial nature of this chapter and its limited length, we acknowledge that we will not be able to cover all of the many excellent works in the field. However, we hope that the discussion here will spark the reader's interest for further exploration of the literature.

11.2 Photonic Crystal Band Structures and Defect Modes

In this section, we give an overview of the fundamental concepts and behavior of photonic crystals. We start by describing the physical origin of the band gap, which arises from coherent scattering of light in periodic materials. Generalizing from the 1D periodic structure of the multilayer film, we go on to discuss 2D periodic photonic crystals, which can block light propagation for any direction in the plane. By introducing defects into the periodic structure, it is possible to create waveguides and microcavities, providing a high degree of control over light propagation. Lastly, we introduce 3D periodic photonic crystals, which can be designed to provide a photonic band gap for arbitrary propagation direction and polarization.

11.2.1 Physical Origin of the Band Gap

To understand the physical origin of the photonic band gap, we can start with a simple, 1D periodic photonic crystal: the well-known multilayer film. A multilayer film (Fig. 11.1(a)) is made up of layers with alternating refractive indices,

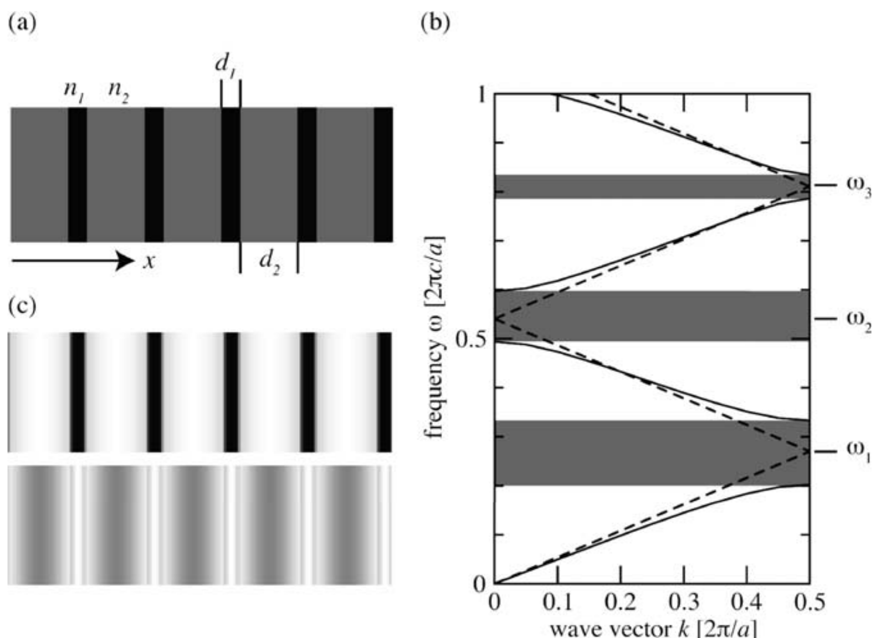


Fig. 11.1 (**a**) Schematic of multilayer film; (**b**) 1D band structure of multilayer film (*solid lines*) and bands of a bulk film with averaged index (*dashed lines*). (**c**) Power in the electric field for the modes of the multilayer film immediately below (*top*) and above (*bottom*) the lowest photonic band gap

$n_1$and n_2. Light propagating through the film reflects from each of the interlayer interfaces. For normal-incidence light, the reflections from each period (bilayer) of the film will interfere constructively provided that the wavelength in air (λ) satisfies the condition

$$m\lambda = 2(n_1 d_1 + n_2 d_2) \tag{11.1}$$

where d_1 and d_2 are the layer thicknesses and m is an arbitrary positive integer. Using the relation $\omega = 2\pi c/\lambda$, we can rewrite this condition as

$$\omega_m = \frac{m\pi c}{n_1 d_1 + n_2 d_2} \tag{11.2}$$

Alternately, we can treat the problem of light propagation in a multilayer film using the language of band structures, or dispersion relations, familiar in solid-state physics. The first step is to rewrite Maxwell's equations in the form of an eigenvalue equation [5]:

$$\nabla \times \frac{1}{\varepsilon} \nabla \times \vec{H} = \left(\frac{\omega}{c}\right) \vec{H} \tag{11.3}$$

along with the constraint

$$\nabla \cdot \vec{H} = 0 \tag{11.4}$$

where $\varepsilon = n^2$ is the position-dependent dielectric function of the material and $\vec{H}$ is the magnetic field. For lossless dielectric functions, Eq. (11.3) is a Hermitian eigenvalue problem with real frequency solutions. Note that we can always obtain the electric field $\vec{E}$ from the solution for $\vec{H}$ from the equation

$$-\frac{\mathrm{i}\omega\varepsilon}{c} \vec{E} = \nabla \times \vec{H} \tag{11.5}$$

For simplicity, we consider an infinite multilayer film. Then due to Bloch's theorem, the solutions of Eq. (11.3), called Bloch waves, take the form of a plane wave times a periodic envelope:

$$\vec{H}(\vec{r}, t) = \mathrm{e}^{\mathrm{i}kx - \mathrm{i}\omega t} \vec{H}_k(x) \tag{11.6}$$

where $\vec{H}_k(x) = \vec{H}_k(x + a)$, and $a = d_1 + d_2$ is the periodicity of the film. In contrast to plane waves, Bloch waves propagate through the crystal *without scattering* – all of the effects of coherent interfacial reflection are accounted for within the Bloch wave form.

As an example, we take $n_1 = 3.45$ and $n_2 = 1.45$, corresponding to the values for silicon and silica at optical communications wavelengths ($\lambda \approx 1.55\ \mu m$), with $d_1 = 0.2a$ and $d_2 = 0.8a$. Equation (11.3) can be solved numerically using the plane-wave expansion method [6]. We plot the lowest few TM-polarized bands as solid lines in Fig. 11.1(b). Here, the TM polarization is defined such that the magnetic field is in the plane of the page, and the electric field is in the perpendicular direction. It is sufficient to plot a finite range of k values between 0 and π/a known as the *irreducible Brillouin zone*, since the dispersion relation is periodic in k with periodicity $2\pi/a$ and symmetric with respect to $k = 0$. Note that frequencies are given in units of $2\pi c/a$, where c is the vacuum speed of light, and wavevector magnitudes are given in units of $2\pi/a$.

In several frequency ranges, indicated by solid gray shading, there are no Bloch wave solutions. These ranges are known as *photonic band gaps*. For frequencies inside a gap, the film will act like a mirror and reflect incident light. This behavior is in agreement with the simple coherent reflection argument given above; the strong reflection frequencies ω_m of Eq. (11.2) fall within photonic band gaps.

To gain further insight into the shape of the photonic band structure, we can look at the multilayer film as a perturbation on a bulk material with an averaged index $\bar{n} = (n_1 d_1 + n_2 d_2)/a = 1.85$. In a bulk material, the solutions to Maxwell's equations are plane waves, and the dispersion relation is given by $\omega(k) = ck/\bar{n}$. In order to plot these solutions in Fig. 11.1(b), we use a mathematical trick. If we consider the bulk material to be periodic with an (artificial) periodicity a, the plane-wave solutions can be rewritten as

$$\vec{H}(\vec{r}, t) = e^{ikx - i\omega t} = e^{i(k - 2\pi m/a)x - i\omega t} e^{i(2\pi m/a)x}$$

where m is an integer chosen such that $k - 2\pi m/a$ falls between 0 and $2\pi/a$, and $e^{i(2\pi m/a)x}$ is the periodic envelope function. The net effect of imposing the artificial periodicity is to "fold" the dispersion relation at the boundaries of the Brillouin zone, as shown by the dashed lines of Fig. 11.1(b).

Perturbing the bulk material to create the multilayer film splits the bands at the folding points, resulting in a gap. The splitting is a consequence of the electromagnetic variational theorem [5]. Due to the perturbation, one Bloch wave tends to concentrate its field in high-n regions, pulling its frequency down. Another Bloch wave is then pushed into the low-n regions, to insure orthogonality with the first. Its frequency is pushed above the bulk value, and a gap results. This is illustrated in Fig. 11.1(c), which shows the electric field energy εE^2 for the modes below and above the first band gap. White corresponds to zero energy, and darker intensities correspond to larger energy values. For the lower-frequency mode, the energy is concentrated in the high-n regions, whereas for the higher-frequency mode, the energy is largely spread out over the low-n region.

Note that in certain frequency regions, the Bloch waves propagate through the multilayer film, or photonic crystal, with dispersion properties quite different from a bulk material. For example, near the band gaps, the slope of the dispersion relation is low, corresponding to slow light speeds (the group velocity $\nu_g = d\omega/dk$).

11.2.2 Two-Dimensional Photonic Crystals

We have shown that the multilayer film, a 1D periodic structure, gives rise to a band gap for propagation in the direction perpendicular to the film layers. To obtain a band gap for any propagation direction in the plane, we can use a structure with 2D periodicity. We will consider the example shown in Fig. 11.2(a), a triangular array of air holes ($n = 1$) in dielectric ($n = 3.45$). Defining a as the center-to-center separation of nearest-neighbor holes, or lattice constant, we choose a hole radius $r = 0.45a$.

As for the multilayer film example, we will plot the solutions to Maxwell's equations in terms of frequency and wavevector. Now, however, we must consider wavevectors in various in-plane directions, which fall within a 2D irreducible Brillouin zone. The irreducible Brillouin zone can be calculated from the basis vectors of the triangular lattice and is shown in Fig. 11.2(b). Figure 11.2(c) shows both the TM and TE dispersion relations. For TE modes, the electric field lies in the plane and the magnetic field is normal to it; for TM modes, the magnetic field lies in the plane. The wavevectors shown on the x-axis run along the outer edge of the irreducible Brillouin zone between the corner points labeled in the inset. In this example, the structure has a *complete photonic band gap* (shaded gray), a frequency region in which there are neither TE nor TM-polarized modes.

In general, careful design is required to achieve a band gap for 2D propagation. Arrays of holes in dielectric tend to favor a TE gap, while arrays of dielectric rods in air tend to favor a TM gap. As for the case shown here, certain structures have a gap for both TE and TM polarizations, while for other structures neither a TE nor a TM gap is present [5].

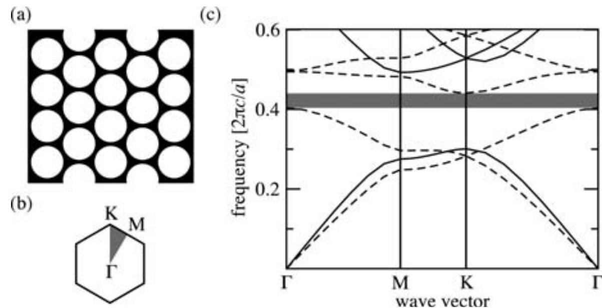


Fig. 11.2 (**a**) Two-dimensional photonic crystal. (**b**) Brillouin zone, with irreducible Brillouin zone shown in *gray*. (**c**) Band structure for TE (*solid lines*) and TM (*dashed line*) modes, with photonic band gap shown in *gray*

11.2.3 Control of Light with Defect Modes

As we have seen, a perfect 2D periodic photonic crystal blocks light propagation for frequencies within the photonic band gap. By deliberately introducing *defects* into the crystal, we can create localized electromagnetic modes that act as waveguides or microcavities.

An example of a waveguide is shown in Fig. 11.3(a). A row of holes within the crystal has been enlarged to have radii of $0.52a$, creating a linear defect. The result is a new solution to Maxwell's equations with a frequency within the photonic band gap. This can be seen from the *projected band structure*, shown in Fig. 11.3(b), which shows the TE modes of the crystal. Modes are plotted as a function of k_x, the wavevector along the waveguide axis. The gray regions indicate *bulk modes*, modes that are spread out throughout the entire 2D photonic crystal. These are similar to the TE modes of Fig. 11.2(c); however, they are now plotted as a function of k_x alone, rather than as a function of a 2D wavevector. The TE band gap extends from 0.30 to 0.49 $[2\pi c/a]$. Inside the gap is a single-mode defect band. Modes in this band are strongly localized near the linear defect region, as shown in Fig. 11.3(c). Intuitively, light is prevented from escaping the defect by the photonic band gap of the surrounding crystal. In contrast to conventional waveguides, which are based on the principle of index guiding, the type of photonic crystal waveguide shown here confines light within a region with lower average refractive index than its surroundings. It is also possible to create a photonic crystal waveguide by increasing the refractive index of a linear defect with respect to its surroundings, for example by decreasing the size of a row of holes or filling them in completely.

An example of a microcavity is shown in Fig. 11.4(a). A single hole has been enlarged to a radius of $0.52a$, resulting in a mode inside the band gap with frequency $\omega = 0.35$ $[2\pi c/a]$. The mode is confined to the defect region (Fig. 11.4(b)) and cannot propagate in the surrounding crystal. Note that for a microcavity mode, it is no longer relevant to plot a band structure. Because the structure including the defect is not periodic in any direction, the solutions to

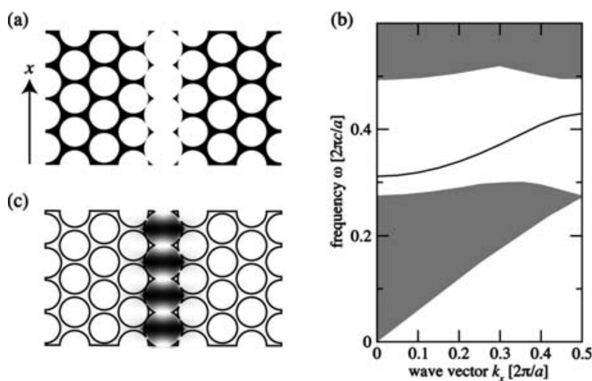


Fig. 11.3 (**a**) Linear waveguide in a 2D photonic crystal created by increasing the radii of a row of holes. (**b**) Dispersion relation for the TE modes of the waveguide. (**c**) Power in the magnetic field for the waveguide mode at $k_x = 0.5$ $[2\pi/a]$

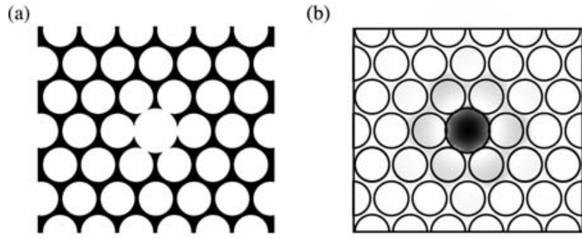


Fig. 11.4 (**a**) Microcavity in a 2D photonic crystal created by increasing the radii of a single hole. (**b**) Power in the magnetic field for the microcavity mode

Maxwell's equations are no longer of the Bloch form and cannot be labeled by a Bloch wavevector.

11.2.4 Three-Dimensional Photonic Crystals

Above, we have reviewed 1D and 2D photonic crystals. It is also possible to design 3D photonic crystals: three-dimensionally periodic structures with a complete band gap or frequency range in which light cannot propagate for any direction or polarization.

Only very particular structures have this property. In general, the crystal must be made up of materials with relatively large difference in refractive index, such as silicon and air, to create strong enough scattering for a complete gap. In addition, the particular geometry must be chosen with care. The face-centered cubic (fcc) lattice, for example, is particularly favorable to the creation of band gaps. Due to its nearly spherical Brillouin zone, the partial band gaps at the corners of the 3D Brillouin zone tend to overlap.

Two examples of 3D photonic crystals are shown in Fig. 11.5. The woodpile structure, shown in Fig. 11.5(a), is made up of stacked layers of parallel rods with square cross-sections. Adjacent layers have perpendicular orientations. The structure has a large photonic band gap of 17% of the midgap frequency for a silicon structure in air [7].

The structure shown in Fig. 11.5(b) is made up of alternating layers of rods and holes. Each layer forms a triangular array. It also has a large photonic band

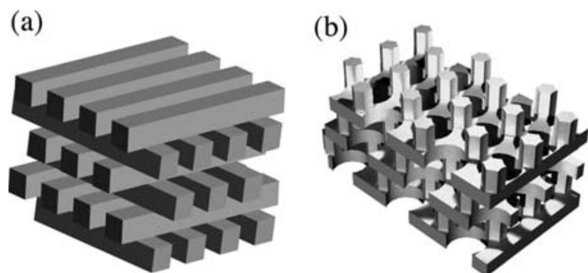


Fig. 11.5 Examples of 3D photonic crystals. (**a**) Woodpile structure. (**b**) Stacked rod and hole layer structure. Both belong to the fcc class of lattices

gap of close to 20% for silicon in air [8]. Because each of the layers resembles a 2D photonic crystal, the structure facilitates the design of waveguides and microcavities based on previously existing 2D designs [9].

Because 3D crystals allow complete confinement of light in three dimensions, they may allow the design of complex, integrated optical circuits with unprecedented control over light flow. However, they are still relatively difficult to fabricate, as will be discussed later in the chapter. For this reason, much experimental research currently focuses on simpler 2D periodic structures known as photonic crystal slabs, discussed in detail in the following section.

11.3 Waveguides and Microcavities in Photonic Crystal Slabs

Photonic crystal slabs are two-dimensionally periodic structures of finite height that approximate many of the useful features of ideal 2D photonic crystals. Their relative ease of fabrication has made them popular for device applications. In this section, we review the basic properties of photonic crystal slabs and describe the design of linear waveguides and microcavities within them.

11.3.1 Band Structures of Photonic Crystal Slabs

An example of a photonic crystal slab is shown in Fig. 11.6(a). The structure is formed by a triangular lattice of holes in a dielectric slab of finite height. Photonic crystal slabs guide light by a combination of two different mechanisms. In the plane, light propagation is similar to that in a 2D photonic crystal. Perpendicular to plane, light is confined by the mechanism of index guiding, since the refractive index of the slab is higher than the surroundings. Modes of

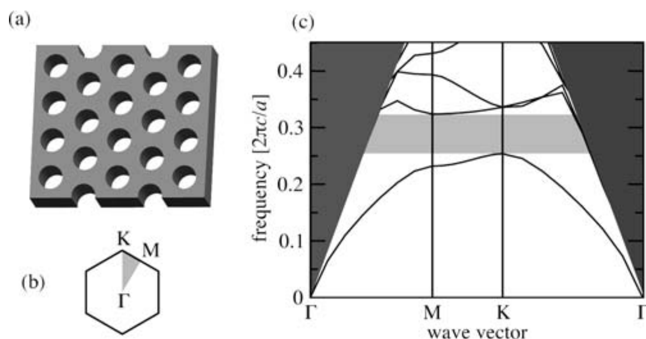


Fig. 11.6 (**a**) Photonic crystal slab. (**b**) Brillouin zone, with irreducible Brillouin zone shown in *gray*. (**c**) Band structure for TE-like modes, with light cone shown in *dark gray* and photonic band gap shown in *light gray*

the slab can be divided into two polarizations, distinguished by their symmetry with respect to the midplane of the slab: even (TE-like) or odd (TM-like) [10].

The Brillouin zone of the photonic crystal slab is shown in Fig. 11.6(b) and resembles that of a 2D photonic crystal. The band structure of the photonic crystal slab for the TE-like polarization is shown in Fig. 11.6(c) for $r = 0.29a$ and slab thickness $h = 0.60a$. A major difference from the band structure of a 2D crystal is the light cone, shown in dark gray. The light cone indicates modes that can propagate in the air above and below the photonic crystal slab. Since the dispersion relation for a plane wave in air is $\omega = ck = \sqrt{k_{\|}^2 + k_{\perp}^2}$, where $k_{\|}$ and $k_{\perp}$ are the magnitudes of the in-plane and out-of-plane wavevectors, respectively, the light cone occupies the region $\omega > ck_{\|}$, where $\vec{k}_{\|}$ lies within the irreducible Brillouin zone of the slab. Modes of the photonic crystal slab that fall in the light cone are not truly guided. Called "leaky modes," they lose light to the surroundings as they propagate. Modes of the photonic crystal slab that lie under the light line are guided in the slab and propagate without loss. Due to the presence of the light cone, there is no complete gap in the band structure. However, there is a gap in the guided modes of the photonic crystal slab, shown in light gray. As we will discuss below, this partial gap can be used to design linear waveguides and microcavities in a similar way as in 2D photonic crystals.

11.3.2 Linear Waveguides in Photonic Crystal Slabs

A waveguide in a photonic crystal slab is shown in Fig. 11.7. In this example, we have chosen to fill in a row of holes. The projected band structure is shown in Fig. 11.7(b). Note that, as for waveguides in 2D photonic crystals, the modes

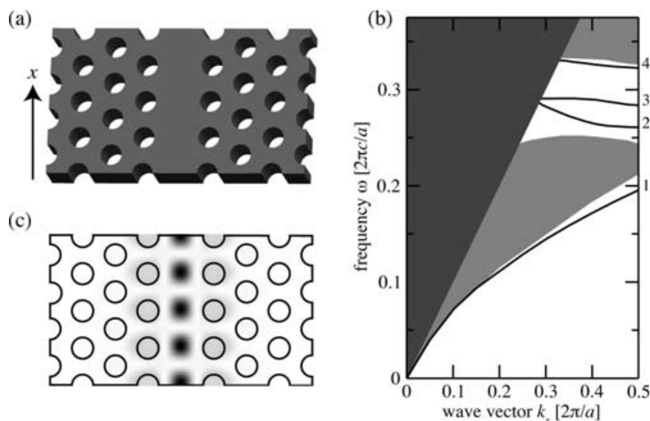


Fig. 11.7 (**a**) Linear waveguide in a photonic crystal slab created by filling a row of holes. (**b**) Dispersion relation for the TE-like modes of the waveguide. (**c**) Power in the magnetic field for waveguide mode 2 at $k_x = 0.5$ $[2\pi/a]$

are plotted as a function of the wavevector along the waveguide axis. The light cone is shown in dark gray and occupies the region $\omega > ck_x$. Modes extended throughout the slab are shown in light gray. Four waveguide bands are also visible, labeled 1 through 4 in the figure. Because the refractive index of the waveguide is higher than its surroundings, it supports an index-guided mode, which lies below the slab modes and is labeled 1. In addition, there are three gap-guided modes, labeled 2, 3, and 4, which are confined to the waveguide region by the photonic band gap in the surrounding crystal. The energy in the magnetic field for band 2 is shown in Fig. 11.7(c).

Many other options exist for designing waveguides in photonic crystal slabs, offering a high degree of flexibility for tailoring the dispersion relation and modal fields. For example, rather than completely filling in a row of holes, the radii of a row of holes can be increased (similar to our 2D example above) or decreased to obtain the desired mode symmetry and frequency [11]. Another option is to surround a strip waveguide with a photonic crystal slab on either side. The result is a large-bandwidth, linearly dispersive waveguide [11,12] with low relative loss [13,14].

Waveguides can also be created in photonic crystal slabs made of finite-height dielectric rods in air. Increasing or decreasing the radii of one or more rows of rods and surrounding a strip waveguide with dielectric rods are all viable means of creating linear waveguides with varying mode profiles and dispersion relations [11].

A different way of making a waveguide is to form a sequence of closely spaced microcavities. In this case, light propagates down the waveguide by tunneling from one microcavity to the next. By increasing the spacing between microcavities, the group velocity can be reduced. Such *coupled-cavity waveguides* [15,16] have attracted great interest in the context of slow light devices for optical delays.

Optical loss is a practical concern for all photonic crystal waveguides. A photonic crystal waveguide mode lying below the light line is ideally lossless. In practice, however, propagation loss results from the scattering of light from small imperfections on the waveguide surfaces. While early experiments measured propagation loss in excess of 10 dB/mm [17], improvements in fabrication accuracy and homogeneity have reduced the loss to 0.6 dB/mm for certain waveguide designs [18]. Due to the short length of typical photonic crystal devices (<1 mm), the total propagation loss can be below 1 dB, an acceptable value for many applications. An additional source of loss is the input coupling to the photonic crystal waveguide from an external light source, such as an optical fiber. Direct coupling from a standard telecommunication single-mode fiber to a photonic crystal waveguide can give rise to loss as high as 30 dB, due to mismatch in waveguide mode profiles and effective indices. However, efficient solutions for minimizing the coupling loss have been developed. For example, a mode converter comprising an in-plane adiabatic inverse taper and polymeric waveguide was experimentally demonstrated to have a low coupling loss of 3–4 dB [18].

Calculation of modes in photonic crystal slab waveguides is far more computationally intensive than for 1D or 2D photonic crystals. Two common calculation methods are the plane-wave expansion method [6] and the finite-difference time domain (FDTD) method [19,20]. For both methods, the memory storage requirements and calculation time increase in proportion to the volume of the problem domain. Recent work has shown that effective index approaches are useful for reducing computational cost. Rather than calculating the full 3D modes of a photonic crystal slab, one first computes the waveguide mode of a solid slab of the same height. The effective index of the slab at a particular frequency of interest is given by $n_{\mathrm{eff}} = ck/\omega$. One then solves a 2D problem with the same geometry as the midplane of the photonic crystal slab, but with a refractive index equal to n_{eff}. Effective index approaches have been shown to have good accuracy for both a modified plane-wave expansion method [21] and the FDTD method [22].

11.3.3 Microcavities in Photonic Crystal Slabs

In an infinite 2D or 3D crystal, microcavities do not suffer from any leakage of light. Leakage is completely prevented by the photonic band gap of the surrounding crystal. In photonic crystal slabs, however, some leakage invariably occurs in the vertical direction, even in an ideal structure free of any structural imperfections. This is because, as pointed out above, introducing a microcavity in a photonic crystal results in an overall structure that is no longer periodic. As a result, modes can no longer be characterized by Bloch wavevectors k, but only by their frequency. A microcavity mode of a given frequency can couple, or leak, to modes in the light cone with the same frequency. To make photonic crystal slab microcavities that are useful for practical applications such as filters and lasers, it is necessary to carefully optimize the structure to minimize such radiation loss.

A key figure of merit for characterizing loss is the cavity quality factor, Q. In the presence of loss, the oscillation of the fields in the cavity is damped in time. Q measures the decay rate of the electromagnetic field energy stored in the cavity in units of the optical period T:

$$Q \equiv 2\pi \frac{\tau_{\mathrm{ph}}}{T} \tag{11.7}$$

where τ_{ph} is the time in which the electromagnetic field energy decays to $1/e$ of its initial value. Assuming exponential decay of the fields in time, we may equivalently write Q in the frequency domain as

$$Q = \frac{\omega_{\mathrm{o}}}{\Delta\omega} \tag{11.8}$$

where ω_o is the center frequency of the resonance and $\Delta\omega = 1/\tau_{ph}$ is the full width at half maximum. Loss lowers Q by broadening the width of the resonance in frequency space. Alternately, Q may be expressed in terms of wavelength as

$$Q = \frac{\lambda_o}{\Delta\lambda} \tag{11.9}$$

where λ_o is the center wavelength and $\Delta\lambda$ is the full width at half maximum.

Naïve microcavity designs, such as filling a single hole in a photonic crystal slab, tend to result in low Q values, of the order of a few hundred or less. One strategy for increasing Q is to delocalize the mode. For example, slightly reducing the radii of a group of adjacent holes [23] represents a weaker perturbation to the underlying crystal, resulting in a more spread-out modal field and a higher Q [24]. In many applications, however, one would prefer to have both a high Q and a small modal volume.

The *multipole-cancellation mechanism* [24] provides one approach to maximizing Q without increasing mode volume. For any microcavity in a photonic crystal slab, it is possible to express the total radiated power as a sum of contributions from different multipole terms. By tuning the structural parameters slightly, it is sometimes possible to cause the lowest-order multipole term (e.g., the electric-dipole radiation term) to vanish, increasing the total Q by several orders of magnitude without pronounced changes in the mode volume.

Alternately, in *the light-cone picture*, the radiated power from the slab may also be expressed as a sum of contributions from outward-going plane waves in air. The amount of power lost to each plane wave can be calculated from a 2D Fourier transform of the near-field of the microcavity mode on a plane above the slab [25]. Only Fourier components with wavevectors lying above the light cone can radiate to air ($k_{\|} < \omega/c$, where $k_{\|}$ is the magnitude of the Fourier wavevector parallel to the slab). Q can be increased by tuning the structure to minimize their value, a process that can be aided by symmetry considerations [26]. The light-cone picture has been used to motivate the design of several types of high-Q microcavities with mode volumes on the order of a cubic wavelength, including modes with dipole [25] and hexapole [27] symmetry. Ultra high-Q microcavities resembling perturbed waveguides [28,29] have been designed with theoretical Q values on the order of 10^7. Experimental Q values of approximately 10^6 have been measured for these structures [30,31], limited by factors such as fabrication imperfections (e.g., sidewall roughness, variation in air hole position, variation in air hole size, etc.).

11.4 Photonic Crystal Surfaces: Surface States, Surface Coupling, Transmission, and Refraction

An infinite photonic crystal has discrete translational symmetry and possesses photonic bands and band gaps as described above. For a semi-infinite photonic crystal, the presence of a surface breaks the translational symmetry normal to

the surface. This results in a spectrum of eigenmodes fundamentally different from an infinite photonic crystal [32]. The most fundamental problem related to a photonic crystal surface is the coupling of an incoming light beam with the eigenmodes of the semi-infinite photonic crystal. These eigenmodes include both surface states (modes) and propagating modes. A rigorous framework for solving this problem is presented in this section. The problem is highly non-trivial, in that a general solution must address both crystallographic surfaces described by integer Miller indices and quasi-periodic surfaces having irrational Miller indices. The transmission theory for the second type of surface is non-existent in solid-state physics, and the theory for the first type has not been discussed in general form. The theory presented here gives a unified rigorous solution to transmission problems through both types of surfaces. The approach is conceptually simple. First, it solves a *surface* eigenmode equation to find all eigenmodes that can be excited by a planar wave impinging upon the photonic crystal; it then solves a set of boundary equations to determine the coupling amplitude for each excited eigenmode.

The importance of this subject in the current research context is related to the recent discovery of the superprism effect and negative refraction, discussed in later sections. The theory presented here should open numerous opportunities for further research in these areas. In a broader context, the theory can be applied to any periodic medium with an ideal "flat" surface, and therefore may supplement our existing knowledge of solid-state physics.

11.4.1 Surface States in a Photonic Band Gap

A finite-sized photonic crystal supports electromagnetic modes that are associated with its surface [5,33]. For simplicity, consider the surface of a semi-infinite 2D square lattice photonic crystal with lattice constant a. The entire system, composed of an air region and a photonic crystal region as shown in Fig. 11.8, has only one dimensional translational symmetry along x. Therefore, such a system is described by two fundamental physical quantities, frequency ω and surface tangential wavevector k_x. Suppose that the original infinite photonic crystal has a photonic crystal band structure described by a function $\omega(k_x, k_y)$. For a given k_x, the set of frequencies corresponding to all possible values of k_y in the first Brillouin zone, $\Omega(k_x) = \{\omega(k_x, k_y)|\ -\pi/a < k_y < \pi/a\}$, covers certain ranges of the spectrum. It is straightforward to see that these and only these ranges contain propagating states of the semi-infinite photonic crystal. On the other hand, states that propagate in air must lie above the light line, $\omega = ck_x$. It is straightforward to classify the modes of the system into four categories: (1) *transmission*: light can propagate both in the PC and air; (2) *external reflection*: light can propagate in air but decays in the photonic crystal; (3) *internal reflection*: light can propagate in the PC but decays in air; (4) *pure surface state*: optical field decays in both regions. These four scenarios are illustrated in Fig. 11.8. Note that the surface states are usually discrete and do not necessarily fill the entire hatched region shown in the

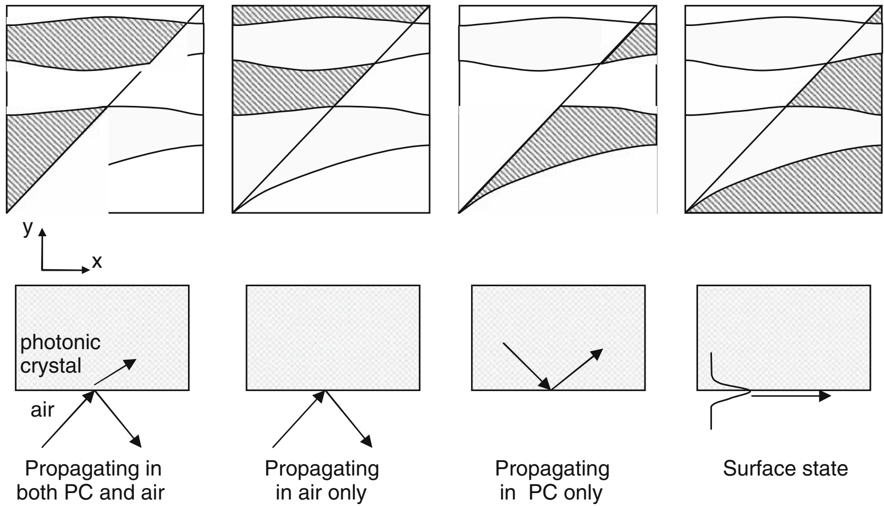


Fig. 11.8 Classification of photonic crystal surface states. *Upper row*: schematic band diagrams. *Lower row*: light coupling on a photonic crystal surface, with *arrows* indicating light transmission/propagation or decay (evanescent wave)

last column of Fig. 11.8. By computing the eigenmodes for a supercell such as the one shown in Fig. 11.9, the states can be found. However, the supercell method has some undesired features. For example, if the surface state has a long decay length into the photonic crystal, then a very large supercell must be used. In addition, the supercell method does not have a general formulation that is applicable to arbitrary surface orientations, which include quasi-periodic surfaces. In the following, we will present a general theory that can handle arbitrary surface orientations for all four scenarios shown in Fig. 11.8.

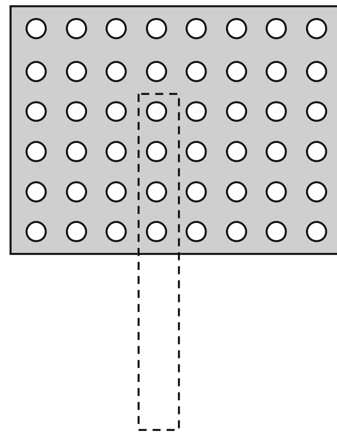


Fig. 11.9 A supercell used for computing surface states

11.4.2 Basic Surface Eigenmode Equations and Fundamental Difference from Bulk Eigenmode Equations

To rigorously study the problem of coupling and/or transmission across a photonic crystal surface, we first derive the basic equations for the modes of a semi-infinite photonic crystal. As we will see, these modes are not entirely included in the set of modes of the infinite photonic crystal, as given by the bulk photonic crystal band structure $\omega(k_x,k_y)$.

We will start with the fundamental electromagnetic equations. For convenience, we consider the TM polarization, whose magnetic field lies in the plane and electric field is normal to the plane. The TE polarization, whose electric field lies in the plane, can be treated similarly. The field equation for the TM polarization is

$$\nabla^2 E(\mathbf{x}) + \frac{\omega^2}{c^2}\varepsilon(\mathbf{x})E(\mathbf{x}) = 0 \tag{11.10}$$

According to Bloch's theorem, we can write $E(\mathbf{x}) = \exp(\mathrm{i}\mathbf{k}\cdot\mathbf{x})\sum_{\mathbf{G}} E(\mathbf{G})$ $\exp(\mathrm{i}\mathbf{G}\cdot\mathbf{x})$. The mode equation has the following form in reciprocal space:

$$-\left[(k_x + G_x)^2 + (k_y + G_y)^2\right]E(\mathbf{G}) + \omega^2\sum_{\mathbf{G}'}\varepsilon(\mathbf{G}-\mathbf{G}')E(\mathbf{G}') = 0 \tag{11.11}$$

This equation, although identical to the equation for photonic band calculations, must be solved in a different way. For photonic band calculations, we solve ω for given values of k_x, k_y. For the surface coupling problem, we solve for the surface-normal wavevector component k_y for given values of ω and k_x.

Consider a finite cutoff of the Fourier series $\mathbf{G} = l\mathbf{b}_1 + m\mathbf{b}_2$, where $l = -L, -L+1, \ldots, L-1, L$, and $m = -M, -M+1, \ldots, M-1, M$. The total number of Fourier components is $N = (2L+1)(2M+1)$. The eigenvalue problem Eq. (11.11) can be written in matrix form as

$$[W][E] = \left(k_y^2[I] + 2k_y[B] + [C]\right)[E] = 0 \tag{11.12}$$

where $[I]$, $[B]$, and $[C]$ are N-by-N matrices, in particular $[I]$ is the identity matrix; and $[E]$ is a N-by-1 column vector whose elements are $E(\mathbf{G}_\mu)$, $\mu = 1, 2, \ldots, N$. The matrix elements are given by

$$B_{\mu\nu} = \delta_{\mu\nu}(\mathbf{G}_\mu)_y \tag{11.13a}$$

$$C_{\mu\nu} = \delta_{\mu\nu}\left[{(\mathbf{G}_\mu)_y}^2 + (k_x + (\mathbf{G}_\mu)_x)^2\right] - \omega^2\varepsilon(\mathbf{G}_{\mu\nu}) \tag{11.13b}$$

In principle, this eigenvalue problem can be solved by calculating the determinant det$[W] = D_W(k_x,k_y,\omega) = 0$. By counting the powers of each variable, one finds that

$$D_W(k_x,k_x,\omega) = \sum_{l=0}^{2N}\sum_{m=0}^{2N}\sum_{n=0}^{2N} c_{lmn}k_x^l k_y^m \omega^n = 0 \tag{11.14}$$

Again, this equation is the same for the photonic band calculation and the surface coupling problem. However, the outcome can be entirely different owing to the known physical quantities in each scenario. Consider a simple (hypothetical) example,

$$D_W(k_x,k_y,\omega) = \omega^2 - (k_x^2 + k_y^2 + k_0^2) = 0$$

where k_0 is a real constant. For the photonic band calculation, we are given the wavevector components k_x, k_y, and we need to find the frequency ω. This equation gives

$$\omega(k_x,k_y) = \sqrt{k_x^2 + k_y^2 + k_0^2}$$

For the surface coupling problem where k_x and ω are known, this equation gives

$$k_y(k_x,\omega) = \sqrt{k_x^2 + k_0^2 - \omega^2} \tag{11.15}$$

Evidently, for real values of k_x, k_y, the $\omega(k_x,k_y)$ is always real. However, for any real values of k_x,ω, the function $k_y(k_x,\omega)$ is not always real. The corresponding eigenmode, $E(\mathbf{G})$, for the surface coupling will become an evanescent mode localized near the surface rather than a propagating mode. It can be proved that the ensemble of eigenmodes for a bulk photonic crystal $\Psi_{\text{bulk}} = \{(k_x,k_y,\omega(k_x,k_y))| -\pi/a<k_x<\pi/a, -\pi/a<k_y<\pi/a\}$ is contained in the ensemble of eigenmodes for a semi-infinite photonic crystal $\Psi_{\text{surf}} = \{(k_x,k_y(k_x,\omega),\omega) \mid -\pi/a<k_x<\pi/a, \omega \in R\}$, where ω runs over the real set R.

It is important to realize that for most practical surface coupling problems with a given pair of k_x,ω, only one or a few eigenmodes have real values of k_y; the set is dominated by evanescent/amplifying modes having complex values of k_y. In other words, very few propagating modes can be excited inside a semi-infinite photonic crystal by a plane wave impinging on its surface. The majority of eigenmodes having complex k_y values do not necessarily lie in a photonic band gap.

11.4.3 Equal Partition of Forward and Backward Eigenmodes

It can be shown that the modes of the semi-infinite crystal system are equally partitioned into forward and backward propagating modes. The beam

propagation direction is determined by the Poynting vector **S**, which is proportional to the group velocity $\mathbf{v}_g$ [32]. For a surface lying along the x-axis, a forward propagating mode must have $S_y>0$, and a backward propagating mode $S_y<0$. Here we introduce an intuitive geometric proof of the equal partition.

The beam propagation direction in a photonic crystal is usually illustrated with the help of dispersion surfaces. A *dispersion surface* is a constant-frequency surface in reciprocal space. A propagating mode having frequency ω and wavevector $\mathbf{k} = k_x\mathbf{e}_x + k_y\mathbf{e}_y$ corresponds to a point located at **k** on the dispersion surface associated with frequency ω. At any point **k**, the group velocity and hence the Poynting vector are parallel to the surface normal of the dispersion surface, $\mathbf{S}||\mathbf{v}_g||\mathbf{n}$. The dispersion surface of an isotropic, homogeneous medium at any given frequency ω is a sphere in 3D, or a circle in 2D as shown in Fig. 11.10(a). The dispersion surface of a photonic crystal may vary in shape as the frequency changes. It can be highly anisotropic as depicted in Fig. 11.10(b). For 2D photonic crystals, the dispersion surface consists of contour lines and may also be called the dispersion contour. For simplicity, we assume that the dispersion contour consists of smooth closed contours. This assumption holds for most practical 2D photonic crystals. The eigenvalue problem of Eq. (11.12) can be solved graphically by intersecting the dispersion contour for the given ω with a vertical line that indicates a constant k_x. Consider two consecutive crossings a, a' on one of the dispersion contours shown in Fig. 11.10(c). From the directions of the outward surface-normal vector **n**, it can be seen that $n_y \propto v_{g,y} = \mathrm{d}\omega/\mathrm{d}k_y$ has opposite propagation directions at a and a'. Indeed, establishing a local polar coordinate system centered at O, we can show that the y-component of the outward surface-normal vector, n_y, always has opposite signs at a and a'. The proof can be applied to any arbitrary pairs such as (b b'), (c c'), and (d d'). A comprehensive proof including the cases of open contours and 3D photonic crystals can be found elsewhere [32].

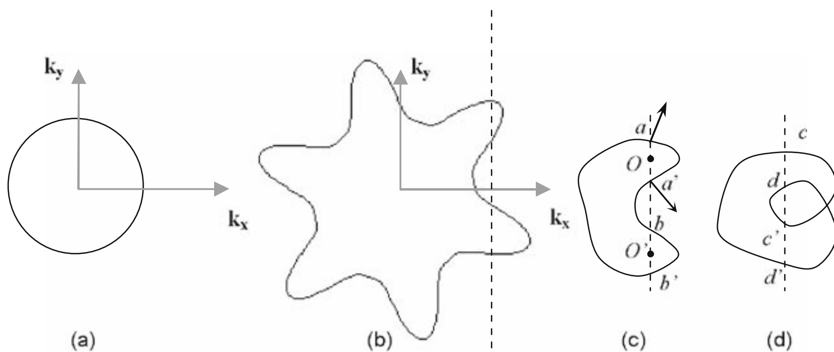


Fig. 11.10 Dispersion contours for (**a**) a homogeneous, isotropic 2D medium; (**b**) a 2D photonic crystal; (**c**, **d**) hypothetical, arbitrary dispersion contours

For a photonic crystal in which $\epsilon(\mathbf{x})$ is a real number everywhere, the complex eigenvalues of k_y always appear in conjugate pairs. Only those with positive imaginary parts give converging modes $\sim\exp(ik_y y)$ in the semi-infinite photonic crystal region, $y>0$.

Only the forward propagating modes and the converging complex k_y modes are physically allowed in the semi-infinite photonic crystal. The preceding analysis indicates that they account for half of the $2N$ eigenmodes found from Eq. (11.12). Following Ref. [32], we call these modes "up modes," forming a set M^+. The other N eigenmodes found from Eq. (11.12) form a set of "down modes," M^-.

11.4.4 Mode Degeneracy and Its Dependence on Surface Orientation

To solve the surface coupling problem, it is necessary to understand the degeneracy of the up and down modes. It turns out that both the up and down modes usually have a high degree of degeneracy, and this degeneracy is closely related to the surface orientation. As we will see later, understanding this degeneracy is necessary for writing a proper set of boundary equations for the coupling amplitudes.

Consider the dispersion contours shown in Fig. 11.11(b) for the (01) surface of a rectangular lattice. Again, we can graphically solve for k_y by locating the intersections of a constant k_x-line with the dispersion contour.

Generally, k_y can take any value between $-\infty$ to $+\infty$. In photonic band calculations, we usually restrict the known quantities k_x and k_y to be in the first Brillouin zone. However, for the surface coupling problem, k_y is initially unknown. Therefore, all intersections shown in Fig. 11.11(b) appear in the set of eigenvalue solutions of Eq. (11.12). Owing to the periodicity in reciprocal space, the subset of these $2M+1$ eigenvalues

$$k_y{}^{(m)} = k_y + mb_2, \quad m = -M, -M+1, \ldots, M-1, M \tag{11.16}$$

is just one "degenerate" solution. It can be proved that the mode fields $E^{(m)}(\mathbf{x})$ are identical in real space. Distinct surface eigenmodes can be restricted to the *1D* Brillouin zone $-b_2/2<k_y<b_2/2$. For complex eigenmodes, the same degeneracy described by Eq. (11.16) holds, and the distinct modes can be located by restricting k_y to the *1D BZ* on the complex k_y plane:

$$-b_2/2 < \mathrm{Re}\, k_y \leq b_2/2 \tag{11.17}$$

Therefore, the number of distinct up modes is

$$N_+ = N/(2M+1) = 2L+1$$

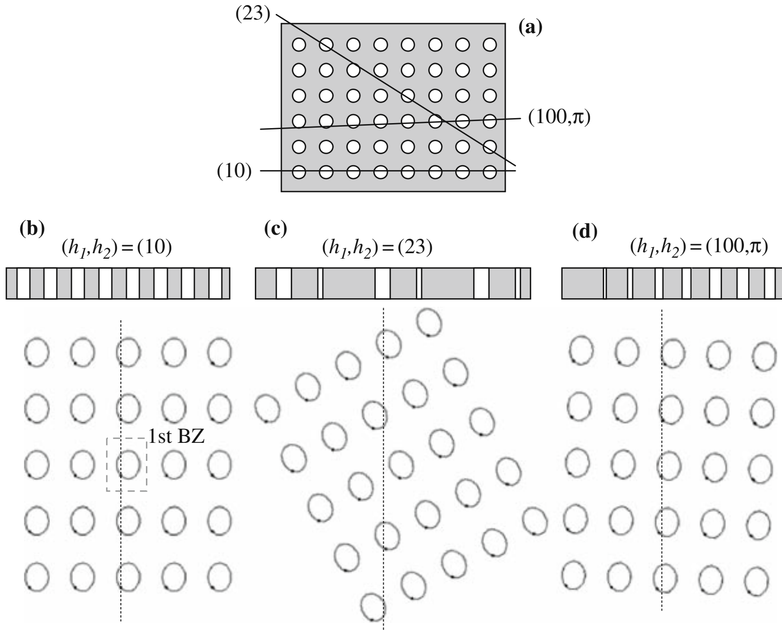


Fig. 11.11 Three types of surface orientations: (**a**) top view of 2D photonic crystal; (**b**–**d**) three surface orientations with side views of the surface in real space and dispersion contours in the reciprocal space (in a 5×5 Brillouin zone)

We note that there is a single transmission amplitude, t_s, for each distinct up mode.

The situation becomes somewhat more complicated for the (11.23) surface depicted in Fig. 11.11(c). For the same drawing showing 5×5 Brillouin zones, the degree of "degeneracy" drops from 5 to 2. An even more complicated situation is depicted in Fig. 11.11(d), where the surface orientation is slightly different from (01). There is no degeneracy in the 5×5 Brillouin zones. The type of surface shown in Fig. 11.11(d) is called a quasi-periodic surface. In both cases, it appears that non-degenerate eigenmodes exist outside the 1D BZ defined by $-b_2/2 < \mathrm{Re}\, k_y \leq b_2/2$ and that the number of t_s is changed from $2L+1$.

A proper treatment of degeneracy involves a prudent choice of the lattice vectors. Consider a surface that has a pair of arbitrary integer Miller indices $(h_1 h_2)$. Without loss of generality, we assume that $0<|h_2|<|h_1|$, and h_2 and h_1 are co-prime. We can redefine the basis vectors for an enlarged lattice cell as

$$\mathbf{A}_1 = h_2\mathbf{a}_1 - h_1\mathbf{a}_2, \quad \mathbf{A}_2 = \mathbf{a}_2 \tag{11.18a}$$

Accordingly, the reciprocal lattice has basis vectors

$$\mathbf{B}_1 = (1/h_2)\mathbf{b}_1, \qquad \mathbf{B}_2(h_1/h_2)\mathbf{b}_1 + \mathbf{b}_2 \tag{11.18b}$$

It is not difficult to show that $\mathbf{A}_1$ always lies parallel to the photonic crystal surface and $\mathbf{B}_2$ is always normal to the surface. Note that Eqs. (11.18a) and (11.18b) return to the original lattice basis vectors for the trivial case $(h_1h_2) = (01)$. If we use this set of new basis vectors in Eq. (11.11) to calculate the Fourier components of the dielectric function and solve for the resulting eigenmodes, then all distinct surface eigenmodes are contained in the 1D BZ defined by $-B_2/2 < \mathrm{Re}\ k_y \leq B_2/2$. The number of distinct up modes remains $2L+1$. For the quasi-periodic surface illustrated in Fig. 11.11(d), the Miller indices must have at least one irrational number. The above procedure does not apply, there is no degeneracy, and hence $N_+ = N$.

11.4.5 Coupling Amplitudes of the Excited Photonic Crystal Modes: Boundary Equations

We can now write the field $E_{\mathrm{I}}(\mathbf{x})$ in the incident medium, and the field $E_{\mathrm{II}}(\mathrm{x})$ in the photonic crystal:

$$E_{\mathrm{I}}(\mathbf{x}) = \exp(\mathrm{i}\mathbf{q}_0\mathbf{x}) + \sum_{\mathbf{G}} r_{\mathbf{p}(\mathbf{G})} \exp[\mathrm{i}\mathbf{p}(\mathbf{G})\mathbf{x}] \tag{11.19a}$$

$$E_{\mathrm{II}}(\mathbf{x}) = \sum_{s\in M^+} \sum_{\mathbf{G}} t_s E_s(\mathbf{G}) \exp\left[\mathrm{i}(k_x + G_x)x + \mathrm{i}(k_y(s) + G_y)y\right] \tag{11.19b}$$

where $r_{\mathbf{p}}(\mathbf{G})$ and t_s are reflection and transmission amplitudes, and the reflection wavevectors are given by

$$\mathbf{p}(\mathbf{G}) = (q_{0x} + G_x)\mathbf{e}_x - \sqrt{\varepsilon_{\mathrm{I}}\omega^2 - (q_{0x} + G_x)^2}\,\mathbf{e}_y$$

The boundary equations can be obtained by matching the $E(x,0)$ and $H_x(x,0)$ for each surface harmonic wave $\exp[\mathrm{i}p_x(\mathbf{G})x]$

$$\delta_{\mathbf{G},0} + r_{\mathbf{p}(\mathbf{G})} = \sum_{s\in M^+} t_s E_s(\mathbf{G}) \tag{11.20a}$$

$$q_{0y}\delta_{\mathbf{G},0} + p_y(\mathbf{G}) r_{\mathbf{p}(\mathbf{G})} = \sum_{s\in M^+} t_s[k_y(s) + G_y]E_s(\mathbf{G}) \tag{11.20b}$$

For a quasi-periodic surface, there are N distinct surface harmonic wavevectors $p_x(\mathbf{G})$, hence the number of boundary equations is $2N$. There are N distinct reflection wavevectors $\mathbf{p}(\mathbf{G})$, and N distinct up modes $k_y(s)$, hence the number of unknowns $r_{\mathbf{p}}(\mathbf{G})$, t_s is $2N$ as well. The boundary equation can be solved.

For a surface that can be described by integer Miller indices, we shall use $\mathbf{G}_{lm} = l\mathbf{B}_1 + m\mathbf{B}_2$ in all subsequent analysis. Because the x-component of $\mathbf{G}_{lm}$ is independent of m, we find $\mathbf{p}(\mathbf{G}_{lm}) = \mathbf{p}(\mathbf{G}_{l0})$. This reduces the number of distinct reflection wavevectors to $2L+1$. The number of distinct surface harmonic wavevectors $p_x(\mathbf{G})$ is also reduced to $2L+1$. Now, let $\mathbf{p}_l = \mathbf{p}(\mathbf{G}_{l0})$. The boundary equations can be rewritten as

$$\delta_{l,0} + r_l = \sum_{s=1}^{2L+1} t_s \sum_m E_s(\mathbf{G}_{lm}) \tag{11.21a}$$

$$q_{0y}\delta_{l,0} + p_{l,y} r_l = \sum_{s=1}^{2L+1} t_s \sum_m [k_y(s) + lB_{1y} + mB_2] E_s(\mathbf{G}_{lm}) \tag{11.21b}$$

Evidently, the total number of unknowns, r_l, t_s, is $2(2L+1)$, and the number of equations is $2(2L+1)$ as well.

Transmission through the (01) surface of a square lattice photonic crystal is shown in Fig. 11.12, where the photonic band gaps are clearly observed. The spectrum calculated with this theory agrees well with the results of the transfer matrix method [34].

If light in an eigenmode of the photonic crystal hits the surface from the inner side, a similar set of boundary equations can be derived for both a periodic surface and a quasi-periodic surface [32].

11.4.6 Some Fundamental Issues of Crystal Refraction or Surface Coupling

Some fundamental aspects of surface refraction/coupling deserve further discussion [32]. Note that most of the following discussion is applicable not only to photonic crystal surfaces, but also to the surface of an arbitrary periodic lattice.

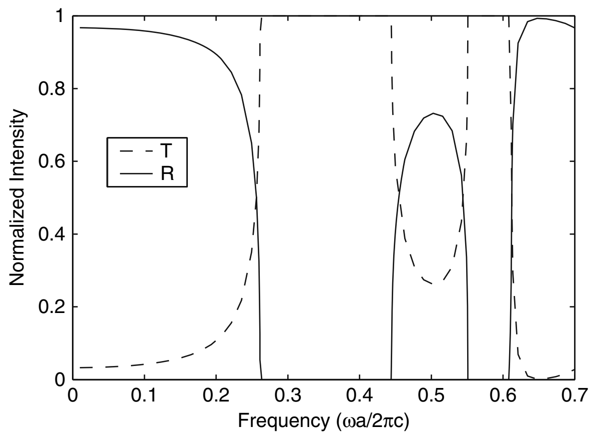


Fig. 11.12 Transmission and reflection spectra for a square lattice (10) surface

Our analysis of PC surface refraction goes beyond merely giving amplitudes r_l, t_s. First, it reveals that incident light recognizes a lattice by its "face." For a square lattice, if the incident surface has Miller indices other than (10) or (01), the light may nevertheless "see" an oblique lattice, according to Eq. (11.18).

For an *ideal* periodic surface, the coincidence of the wavevectors of different reflection orders, $\mathbf{p}(\mathbf{G}_{lm}) = \mathbf{p}(\mathbf{G}_{l0})$, means that the reflected waves carry only the information of the surface periodicity. In a sense, the other Bragg planes inside the PC are hidden. Simply from the reflected waves, one could not tell whether the crystal is a 1D grating or a 2D photonic crystal. It is interesting to notice that the degeneracy of the PC modes is governed by the *surface-normal 1D BZ* associated with $\mathbf{B}_2$, whereas the wavevector difference of the reflected waves is dictated by the *surface BZ* through the new surface wavevector $\mathbf{B}_1$. Although $\mathbf{B}_1$ is generally not parallel to the surface, its x-component enters the reflection wavevector difference as

$$p_{l,x} - p_{l-1,x} = B_{1x} = 2\pi/A_1 \tag{11.22}$$

It has been noted in previous studies of (10), (100), or (111) surfaces that the reflected modes carry only the information of the surface periodicity in their wavevectors (for example, in Refs. [34,35]). However, such a phenomenon has not previously been correlated with the mode degeneracy of the photonic crystal.

For a quasi-periodic surface, the reflected waves carry the information of all Bragg planes, not just the surface BZ. This is obvious because no two $\mathbf{p}(\mathbf{G}_{lm})$, $\mathbf{p}(\mathbf{G}_{lm'})$ coincide, and for a quasi-periodic surface, there is actually no surface BZ owing to the lack of surface periodicity.

Another implication of the theory is that a slight change of surface orientation may split one PC beam into many beams. Consider the rectangular lattice example we discussed above. For the (01) surface, suppose there is only one propagating mode among $2L+1$ distinct up modes. One can easily show that when L increases, the new modes introduced will all be evanescent modes with complex $k_y(s)$; the change of M does not affect the number of distinct up modes. On the other hand, it is a drastically different case for a quasi-periodic surface that could be merely 0.0001 degree from the (01) direction. Owing to the lack of degeneracy, the number of distinct up modes will increase when M increases. Particularly, the number of distinct propagating modes will increase as M, which means more beams will be present in the crystal. How to observe such a sensitive phenomenon is an interesting question. Note that a quasi-periodic section of acceptable quality cannot be achieved in an atomic crystal because an atom cannot be divided or "cut" into fractions. Whereas artificial structures such as photonic crystals can form an ideal "flat" surface, an atomic crystal surface intended to be a quasi-periodic section would in general appear ragged or have lattice voids.

Furthermore, to calculate the photonic bands of 2D and 3D photonic crystals, one has an infinite number of choices of basis vectors of the unit cell giving equivalent results. However, the choices become limited in the refraction problem as seen in Eq. (11.18a). From a symmetry point of view, the presence of the surface breaks discrete translational symmetry of a crystal. This causes the 2D periodicities in the real and reciprocal spaces to be sectioned along the x and k_y directions, respectively. For a crystallographic surface described by integer (rational) Miller indices, the 2D translational symmetry is decomposed into the 1D translational symmetry along the x-axis in the real space and the 1D "degeneracy" of k_y in the reciprocal space. The lower translational symmetry limits the choices of primitive translation vectors to a subset of those of an infinite crystal. Specifically, one of the basis vectors must be a multiple of $\mathbf{A}_1$ defined in Eq. (11.18a) so as to reflect the surface periodicity. The other basis vector can be arbitrarily chosen.

The theory presented above is valid for any surface termination. Note that as the surface termination changes, the Fourier transform $\epsilon(\mathbf{G})$ *implicitly* gains a phase factor, $\exp[iG \cdot (\Delta y \mathbf{e_y})]$, where Δy measures the y-shift of the surface with respect to one cell center. Therefore, the surface termination information enters the field equation implicitly through $\epsilon(\mathbf{G})$. Surface termination can also be treated explicitly. In the next section, we will see that the boundary conditions in Eq. (11.24) involve phase factors $\exp[ik_y(s)d]$ and $\exp[i\mathbf{G_{lm}}(d\mathbf{e_y})]$ for the back surface of a photonic crystal slab. When explicitly treating the surface termination in the semi-infinite crystal problem, similar factors will appear in Eqs. (11.21a) and (11.21b). Lastly, the surface is terminated at the centers of air holes for the triangular photonic crystal problem treated in Ref. [32], whereas the termination is on the middle plane between two cylinders for the square lattice problem whose results are plotted in Fig. 11.12.

The surface transmission and coupling theory given here can be extended to 3D photonic crystals. The details can be found in a recent publication [36].

The theory presented in this section can be readily extended to treat the surface coupling/transmission problem for any matter wave and any semi-infinite periodic medium as long as the wave equation is linear. It is our understanding that some problems studied in this section, though fundamental to solid-state physics, have not been systematically investigated as done here. Further application of this theory to other surface problems may illuminate some unexplored aspects of solid-state physics.

11.5 Transmission Through an Arbitrary Photonic Crystal

A number of numerical and theoretical methods have been employed to study light transmission (or scattering) through photonic crystals, including the transfer matrix method [34,37], the scattering theory of dielectric cylinder/

sphere lattices [35,38,39] or multiple scattering method [40], and the internal field expansion method [41,42]. In certain scattering problems, a photonic crystal is comparable to or smaller than the light beam cross-section, and we are interested in the far-field properties of the scattered light. Such a scenario is relevant, for example, to spectrum measurement and investigations of optical loss in certain photonic crystal structures. However, for most integrated photonic devices, we are more interested in transmission problems that deal with the optical field inside a photonic crystal or near its surface. This section will focus on theories that can effectively handle these problems.

11.5.1 Transmission Theory for a Photonic Crystal Slab

A number of methods have been developed for computing the transmission through finite-sized photonic crystals. Pendry and MacKinnon developed a general computational method for calculating photonic crystal transmission using the transfer matrix approach [37], which has been validated using microwave photonic crystal measurements [3]. For photonic crystals composed of spheres or cylinders, scattering theories can be employed to compute the transmission and reflection coefficients [38,39]. For structures with piecewise constant refractive index, mode matching techniques with perfectly matched layer boundary conditions can be used [43]. Sakoda has developed an internal field method to compute the transmission [41,42].

These methods can be classified into three categories, according to their effective computational domains. The first category is the whole-space methods, which need to compute the electromagnetic field in the entire space (or in the entire photonic crystal). The finite-difference time domain technique is a typical whole-space method. The second category is the 1D supercell methods, which compute the field in a 1D supercell spanning from the entrance surface to the exit surface. Such a method is applicable only if the photonic crystal is a slab with parallel front and back surfaces. The internal field expansion technique [42] is an example for this category. The third category is the single-cell method, where we need to compute the field in only one cell per surface to determine the transmission through the entire photonic crystal [32,34]. Evidently, the single-cell methods are the most efficient due to their small computational domain.

Here, we extend the surface transmission and coupling theory presented in the preceding section to compute the transmission through a photonic crystal slab [44]. Note this approach falls into the single-cell method category. Essentially, the slab transmission problem requires that the boundary equations at the front and back surfaces be solved simultaneously. Similar to the single-surface transmission problem, the fields in front of, inside, and behind the photonic crystal slab can be expressed as

$$
\begin{aligned}
E_{\mathrm{I}}(\mathbf{x}) &= \exp(\mathrm{i}\mathbf{q}_0\mathbf{x}) + \sum_l r_l \exp(\mathrm{i}\mathbf{p}_l\mathbf{x}) \\
E_{\mathrm{II}}(\mathbf{x}) &= \sum_s \sum_{l,m} c_s E_s(\mathbf{G}_{lm}) \exp\left[\mathrm{i}(k_x + lb_1)x + \mathrm{i}(k_y(s) + mb_2)y\right] \\
E_{\mathrm{III}}(\mathbf{x}) &= \sum_l t_l \exp(\mathrm{i}\mathbf{v}_l\mathbf{x})
\end{aligned}
\tag{11.23}
$$

where the reflection and transmission wavevectors are given by

$$
\begin{aligned}
\mathbf{p}_l &= (q_{0x} + lb_1)\mathbf{e}_x - \sqrt{\varepsilon_{\mathrm{I}}\omega^2 - (q_{0x} + lb_1)^2}\,\mathbf{e}_y \\
\mathbf{v}_l &= (q_{0x} + lb_1)\mathbf{e}_x + \sqrt{\varepsilon_{\mathrm{III}}\omega^2 - (q_{0x} + lb_1)^2}\,\mathbf{e}_y
\end{aligned}
$$

and c_s, r_l, and t_l represent the complex amplitudes of the eigenmode $E_s(\mathbf{x})$, the lth order reflected wave, and the lth order transmitted wave, respectively. Note $q_{0x} \equiv k_x$. The boundary conditions for the front and back surfaces are given as

$$
\begin{aligned}
\delta_{l,0} + r_l &= \sum_s c_s \sum_m E_s(\mathbf{G}_{lm}) \\
q_{0y}\delta_{l,0} + p_{l,y} r_l &= \sum_s c_s \sum_m [k_y(s) + mb_2] E_s(\mathbf{G}_{lm}) \\
\mathrm{e}^{\mathrm{i}v_{l,y}d} t_l &= \sum_s c_s \mathrm{e}^{\mathrm{i}k_y(s)d} \sum_m E_s(\mathbf{G}_{lm}) \exp[\mathrm{i}\mathbf{G}_{lm} \cdot (d\mathbf{e}_{\mathbf{y}})] \\
v_{l,y}\mathrm{e}^{\mathrm{i}v_{l,y}d} t_l &= \sum_s c_s \mathrm{e}^{\mathrm{i}k_y(s)d} \sum_m [k_y(s) + mb_2] E_s(\mathbf{G}_{lm}) \exp[\mathrm{i}\mathbf{G}_{lm} \cdot (d\mathbf{e}_{\mathbf{y}})]
\end{aligned}
\tag{11.24}
$$

For illustrative purposes, we apply our theory to a grating diffraction problem. Note that the surface-relief grating shown in Fig. 11.13(a) can be regarded as a monolayer of a 2D photonic crystal as shown in Fig. 11.13(b).

The Fourier components of the sinusoidal grating are given by

$$
\varepsilon(\mathbf{G}_{lm}) \begin{cases} = \frac{1}{2}(\varepsilon_{\mathrm{I}} + \varepsilon_{\mathrm{III}})\delta_{l,0} + \frac{1}{4}(\varepsilon_{\mathrm{III}} - \varepsilon_{\mathrm{I}})(\delta_{l,1} + \delta_{l,-1}) & m = 0 \\ \dfrac{\varepsilon_{\mathrm{III}} - \varepsilon_{\mathrm{I}}}{2\pi m}\left[\mathrm{i}\delta_{l,0} + (-1)^m \mathrm{i}^{l-1} J_l(m\pi)\right] & m \neq 0 \end{cases}
\tag{11.25}
$$

The transmission and reflection coefficients, known as diffraction efficiencies in grating terminology, are plotted in Fig. 11.13(c). The results obtained from the theory are in good agreement with the rigorous coupled wave approach [45].

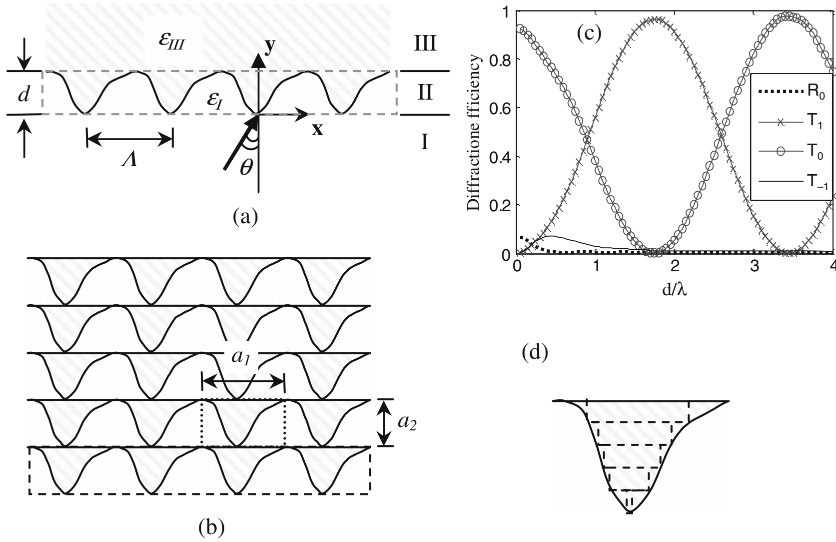


Fig. 11.13 Grating as a monolayer photonic crystal [44]. (**a**) Grating structure; (**b**) virtual photonic crystal; (**c**) diffraction efficiencies; (**d**) illustration of the staircase approximation (or slicing) used in other approaches

We should mention that most theories assume that the incident light is a plane wave of infinite lateral extent for the slab transmission problem. Therefore, the internally reflected waves from the front and back surfaces fully overlap inside the slab, producing the interference effect seen in the spectral oscillation in Fig. 11.6. If the incident beam is sufficiently narrow and the slab is thick, then different scenarios could occur. Consider a simple situation where each internal reflection generates only a single reflected beam inside the PC slab. After a round trip of reflections, the beam will be shifted laterally with respect to the original beam in the PC slab. If the reflection angles are relatively large and/or the beam widths are relatively narrow, the secondary beams generated due to multiple reflections may have little spatial overlap with the original beam. The field immediately outside each surface of the slab will consist of an array of parallel beams rather than a single beam that contains the interference effect. The slab transmission theories for the planar incident waves cannot predict the strength of each exiting beam in such a case. Instead, the single-surface refraction theory developed here must be used.

Moreover, there are many practically valuable cases where the entry and exit surfaces are not parallel to each other [46]. Study of the single-surface transmission problem is necessary to understand these diverse situations, which frequently arise in the design of useful devices [46]. Note that the single-surface transmission problem for a photonic crystal is nothing but refraction. Moreover, a PC slab transmission theory can be obtained from a refraction theory (as

is done in conventional thin film interference theory), but not the reverse. As Li and Ho pointed out [34], for a planar incident wave (of infinite beam width), the mathematical solution for the internal field deep in an extremely thick slab does not converge to the true solution of the field in a semi-infinite photonic crystal, because multiple reflections always exist in the slab, regardless of the separation between its surfaces.

11.6 Fabrication of Photonic Crystals

The earliest attempts to fabricate photonic crystals began with structures with relatively large feature sizes. The wavelength of a photonic band gap scales with feature dimensions: empirically, the band gap appears at a wavelength around three times the lattice constant. Therefore, a periodic structure on the millimeter scale provides a band gap in the microwave part of the electromagnetic spectrum. Soon after proposing the photonic band gap concept, Yablonovitch began to fabricate structures to test his ideas. He proposed to drill three series of holes into a dielectric material along three crystallographic axes, the (110), (101), and (011) directions of a diamond lattice. This structure, which has a complete band gap and is amenable to common machining methods at microwave frequencies [47], was later called Yablonovite. In the mid-1990 s, with the advance of electron-beam (e-beam) lithography, researchers commenced the effort to fabricate photonic crystals for optical wavelengths. A variety of fabrication methods have been developed. In this section, we will briefly review a number of the most common methods to lay a foundation for our subsequent discussion of photonic crystal devices.

11.6.1 Electron-Beam Lithography

Today, electron-beam nanolithography facilities capable of patterning feature sizes around 50 nm or smaller are widely available in academic and industrial research laboratories. The resolution of this lithography technique is sufficient for patterning photonic crystals for most optical and infrared wavelengths.

The technique starts with a relatively flat piece of material, for example, a silicon or GaAs wafer. A layer of e-beam resist, an analog of the photoresist used for photolithography, is spin-coated on the wafer surface. The most common e-beam resist is PMMA, although other types of resists are also used. The wafer is then loaded into the e-beam chamber for patterning. Computer software is used to control the electron beam, scanning it over the surface of the wafer to write the desired pattern. For a positive resist such as PMMA, the exposed areas are dissolved in the subsequent developing process. For a negative resist, the exposed areas of the resist remain on the wafer whereas the

unexposed areas are dissolved. The pattern is then transferred on to the underlying silicon or GaAs by wet or dry etching. Dry etching, particularly reactive ion etching, is often preferred so as to produce a vertical side wall in the etched regions of the wafer. As the e-beam resist itself is not sturdy enough to sustain extended dry etching times, a thin intermediate layer, for example silicon oxide, may be grown on the wafer surface before coating the e-beam resist. The exposed resist is used as a mask to etch the thin layer of oxide in a time period short enough not to compromise the e-beam resist. Once the pattern is transferred from the resist to the oxide, one can etch the underlying silicon using the patterned oxide layer as a "hard" mask. With a proper dry etching recipe, silicon can be etched significantly faster than silicon oxide.

Krauss et al. first patterned a 2D photonic crystal slab on an AlGaAs substrate using e-beam lithography [48]. Transmission, reflection, and diffraction of 2D photonic crystal slabs at near-infrared wavelengths were subsequently measured quantitatively [49]. In another notable advance, Lin and Ho fabricated a 3D woodpile photonic crystal with a full photonic band gap [50].

11.6.2 Holography

Holographic methods provide another way of fabricating photonic crystals. A holographic pattern is formed by the interference of coherent beams to produce a standing wave in 3D space. For N beams, the spatial optical intensity may be written as

$$I \sim \left| \sum_{m=1}^{N} \mathbf{E}_m \cos(\mathbf{G}_m \cdot \mathbf{x} - \omega t + \phi_m) \right|^2 \tag{11.26}$$

The modulation of the optical intensity is given by the cross terms:

$$I \sim \sum_{l,m} \mathbf{E}_l \mathbf{E}_m \cos[(\mathbf{G}_l - \mathbf{G}_m) \cdot \mathbf{x} + (\phi_l - \phi_m)]$$

There are $N(N–1)/2$ non-zero spatial modulation wavevectors $\mathbf{G}_{lm}$. For $N = 3$, there are three spatial modulation wavevectors lying in one plane, therefore they can only form a 2D photonic crystal. At least $N = 4$ beams are required to produce a 3D photonic crystal. It has been shown that all five 2D and fourteen 3D Bravais lattices can be constructed via holography [51].

The spatially modulated light intensity is used to expose a photosensitive polymer. Since the difference in refractive index between exposed and unexposed regions is quite small (<0.1), it is necessary to dissolve away regions of the pattern, leaving air voids in the polymer. For a positive resist, regions that

receive an exposure dosage above a certain threshold value will be removed in the subsequent developing process. For a negative resist, the less exposed regions are removed. The remaining structure has a relatively large index contrast, $n_{polymer}:n_{air}\sim 1.5:1$, sufficient to form a wide band gap along certain crystallographic directions.

With currently available high-power, high coherence length lasers, it is easy to enlarge the beam beyond a few centimeters and produce large area 2D or 3D holographic photonic crystals. This essentially parallel exposure process may form a photonic crystal in a few seconds, significantly faster than the e-beam lithography technique. However, unlike e-beam lithography, holographic lithography is limited to periodic patterns and cannot produce waveguide or microcavity structures. The recent advance of prism holography [52] has significantly reduced the cost of producing a large number of coherent beams. However, there remain challenges to fabricating thick 3D photonic crystals by holography.

First, the exposure process requires a material that absorbs light. As a result, the intensity of the beams decreases with distance into the photosensitive polymer, and the bottom of the sample (furthest from the light source) tends to be underexposed. In addition, the top of the sample reacts with the developer for a longer time than the bottom, since the developer must dissolve away voids at the top of the crystal to reach the region below. For a positive resist, these effects combine to give severe asymmetry for a thick 3D crystal. One way to overcome the problem is to expose from the backside of the substrate [53]. For a negative resist, the two effects tend to cancel each other out in the standard front-surface exposure setup. For both positive and negative resists, the reduction of optical intensity and effective developer concentration in the deep body of a thick film limits the maximum thickness of a holographic photonic crystal. Further study is needed to overcome this limitation.

Like all optical lithography techniques, the holographic method also suffers from the drawback that the minimum resolution is limited to a half of the wavelength. Nonetheless, with common blue or violet lasers as coherent sources, the resolution is adequate for a wide range of applications at 1300 and 1550 nm wavelengths.

Interestingly, the study of photonic crystals has brought a new perspective to research on holographic structures. While past research on holograms was focused on their far-field diffractive properties, research on holographic photonic crystals focuses on how the band structure affects light propagation inside the material or near its surface.

11.6.3 Laser Direct Writing by Two-Photon Absorption

Recently, laser direct writing techniques have been developed to fabricate moderate-sized 3D photonic crystals. A photosensitive material is coated on a substrate, and a laser beam is scanned through the body of the photosensitive

material to expose certain areas. Two-photon photopolymerization (2 PP) has been widely employed to achieve highly localized structural features (see Chapter 12 in this book). Using a lens, the laser beam is focused on a spot. Polymerization occurs only where the laser energy density exceeds a threshold value, which occurs in a small 3D volume near the focal point. Since the threshold for 2 PP is generally higher than for standard single-photon polymerization, the resolution of the method is finer.

The resolution can be further improved by using pulsed lasers. For a laser beam (or focal spot) with a beam waist d_0 and fluence F, the diameter of the polymerization region is [54]

$$d = d_0 \sqrt{\ln(F/F_{\mathrm{th}})}$$

where F_{th} is the threshold fluence for polymerization. Using a femtosecond pulsed laser, it is possible to precisely control the dose so that F is only slightly above threshold. According to the above equation, the polymerization diameter d is then much smaller than the laser beam/spot diameter d_0. In principle, the resolution can be arbitrarily small. In practice, however, the beam quality, beam profile stability, and intrinsic fluctuation of the laser field limit the minimum resolution to approximately 100 nm.

One limitation of the method is that during the serial exposure process, the top region of the photopolymer absorbs a small dose of laser energy while the focus of the beam scans through the lower portion of the film. This partial dosage must be compensated to fabricate a thick 3D photonic crystal. While stereography techniques [55,56] have successfully produced 3D structures with cross-sectional areas exceeding $100 \times 100\ \mu\mathrm{m}^2$, patterning of thick 3D photonic crystals with submicron feature sizes has yet to be demonstrated.

11.6.4 Self-Assembly and Templating

Identical particles of certain materials (e.g., submicron polystyrene spheres) dispersed in a liquid have the tendency to self-assemble into a crystalline phase, usually a stack of hexagonal close-packed planes. If the stacking is properly ordered, the result is a face-centered cubic structure. While such a colloidal crystal is mechanically unstable, the voids between the spheres can be filled with another material to form a solid structure. Liquids or gases are used to carry the fill material through the micro-channels between the spheres. Furthermore, one may remove the original lattice of microspheres by sintering or chemical reaction, forming an inverse of the original structure (Fig. 11.14). The original method for forming colloidal photonic crystals employed equally sized emulsion droplets as the template to form a periodic macroporous material [57].

Latex [58], polystyrene [59], silica [60], and metal [61] have been used as microsphere materials. A variety of materials, such as silicon oxide, alumina,

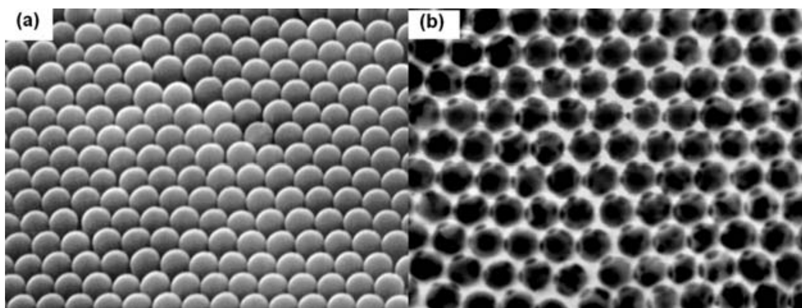


Fig. 11.14 (a) Colloidal photonic crystal template; (b) Structured porous silica made using the template [62]. (original high resolution micrographs courtesy of Orlin D. Velev)

titania, and CdSe have been used to fill the interstitial voids, as discussed in a review [62]. Vlasov et al. have developed a low-pressure chemical vapor deposition method to fill the voids with silicon [60], a desirable material for integrated photonics applications. The microspheres are then removed by wet etching. The resulting structure has a complete band gap of 10% of the midgap frequency.

It is also possible to form structures other than the close-packed hexagonal planar stacks or fcc lattices. Surface preparation prior to self-assembly is generally required to obtain repeatable results [63].

In principle, the self-assembly method is a high-throughput approach that is capable of patterning large 2D or 3D photonic crystals in a short time. However, the method is subject to various types of defects, such as stacking faults, dislocations, and point defects. The quality of a colloidal crystal is also affected by the size distribution (or mass dispersion) of the microspheres. Thicker photonic crystals are often desired, but the quality and optical properties of 3D colloidal photonic crystals may vary with their thickness [64]. Further development is needed to overcome these problems and demonstrate low defect-density, large area, thick 3D photonic crystals in a preferred photonic material.

11.6.5 Nanoimprint

In nanoimprint lithography, a mask pattern is defined by pressing a template, or mold, against a resist layer on the surface of a substrate. The imprinted resist is then used to etch the pattern into the substrate. In some nanoimprint methods, the mask is made of a flexible polymer material, such nanoimprint techniques are sometimes called “soft-molding,” in contrast to “hard-molding” techniques that use rigid templates made out of inorganic materials such as silica. One advantage of nanoimprint lithography is that once the template is made (using e-beam lithography or other techniques), it can be reused repeatedly, yielding a relatively fast, low-cost, high-volume patterning technique.

As an example, we consider the patterning of a polymeric 2D photonic crystal using a polydimethylsiloxane (PDMS) template [65]. The nanoimprint process starts with a master structure on the e-beam resist ZEP520A directly patterned by e-beam lithography. The PDMS prepolymer in 10:1 mixing ratio is then poured onto the master structure of baked e-beam resist patterns. After being cured at 60°C for 12 h, the PDMS is peeled off from the e-beam resist, and a PDMS template (soft mold) is obtained on the PDMS film. A thin film of ultraviolet curable photopolymer is coated on a substrate, and the PDMS template is placed in contact with the photopolymer layer from the top. In some embossing processes employing a hard mold, pressure may be applied to the template to imprint the pattern. In this case, the capillary force drives the uncured polymer solution to fill the recesses of the template, leading to pattern formation. With a low-viscosity polymer, it is possible to fill the voids in microseconds. As the PDMS template (a few millimeters thick) is largely transparent, ultraviolet irradiation through the PDMS is employed to cure the polymer. A patterned 2D polymeric photonic crystal is shown in Fig. 11.15.

Due to the limited time and space, we limit our discussion to one example on this topic. The readers are encouraged to explore the other nanoimprint references contained in Ref. [65].

11.6.6 Other Techniques

A variety of other methods have been employed to fabricate photonic crystals. In a joint effort of Sandia Laboratories and Iowa State University, a 3D woodpile

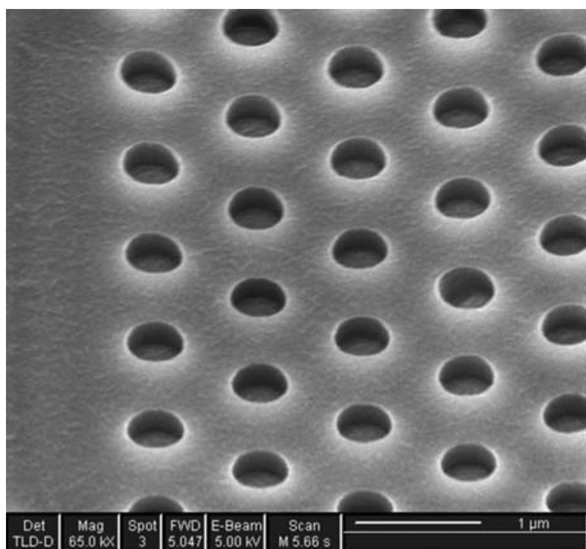


Fig. 11.15 A 2D polymeric photonic crystal imprinted by a PDMS template

photonic crystal was fabricated using standard microelectronics fabrication techniques. A layer of SiO_2 was deposited, patterned, and etched to form the shafts of SiO_2. The resulting trenches were filled with poly-silicon, and the surface was planarized using chemical mechanical polishing. A similar sequence of processes was repeated to form the second layer, and so on. Finally, the silicon oxide was removed by dipping into an HF solution, and a four-layer 3D photonic crystal made of poly-Si was formed. A 3D crystal containing microcavities has been made in a similar fashion [66]. In other work, a Germanium inverse woodpile structure has been fabricated using a polymer template [67].

Another method first patterns and dry etches a 2D hexagonal lattice on a silicon or silica substrate. Alternating deposition of silicon and SiO_2 preserves the lattice pattern/topology of the bottommost layer, forming a graphite-like 3D photonic crystal composed of up to 20 pairs of Si/SiO_2 layers [68].

The wafer fusion technique was employed to fabricate 3D photonic crystals in GaAs or InP [69]. A pair of striped compound semiconductor wafers was first fused at around 700°C. One of the substrates was then removed by a combination of chemical and dry etching. Two such wafers, each having two layers of stripes, were again wafer-fused to form a four-layer structure, using a laser-assisted alignment technique. Such a process can be continued to form close to 20 layers. Point defects and line defects can also be introduced by patterning one of the layers and fusing it into the 3D photonic crystal.

11.6.7 Process Integration

As photonic crystal research evolves from science toward technology, increasing emphasis is being placed on making functional photonic crystal devices that can be integrated with other photonic and electronic structures and devices. To this end, it is important that methods of fabricating photonic crystal structures are compatible with the methods of fabricating accompanying photonic/electronic structures on the same substrate. As an example, we consider the lithography technique used in patterning photonic crystals. Although e-beam lithography is most widely employed in academic research, deep ultraviolet (DUV) photolithography actually suffices for patterning 2D photonic crystals for telecommunication wavelengths, which require a lattice constant around 400 nm for silicon and common III–V materials. For integrating photonic crystal structures with electronics on a single chip, DUV or even shorter wavelength photolithography tools are likely to be used in mass production, due to cost and throughput concerns. In this regard, Krauss's group in the UK and Intel employed 193 nm DUV lithography to pattern and fabricate 2D photonic crystals with lattice constants as fine as $a = 280$ nm [70]. A fabricated photonic crystal waveguide is shown in Fig. 11.16.

Many 2D photonic crystal structures have the topology of a membrane, i.e., a free-standing photonic crystal slab in air. Such a structure is usually

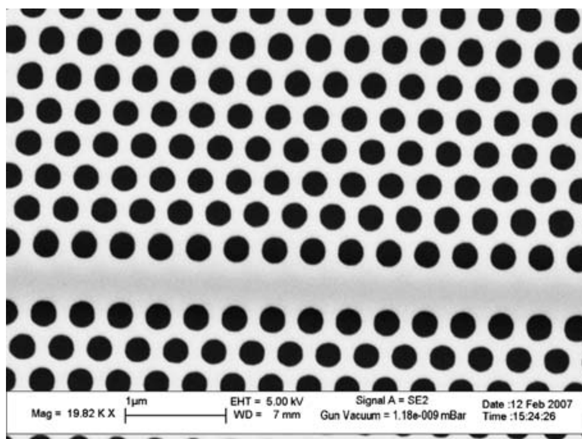


Fig. 11.16 Photonic crystal structures made by DUV lithography (courtesy of Thomas F. Krauss)

formed by wet etching part of the substrate underneath the photonic crystal slab. For a simple photonic crystal structure, we may pattern large openings surrounding the photonic crystal region to deliver a sufficient amount of etchant to the substrate region directly underneath the photonic crystal area [71]. However, a realistic, complex photonic crystal with electronic structures may not allow such latitude. For example, in one recently demonstrated photonic crystal laser structure [72], the photonic crystal defect cavity has a semiconductor post of submicron diameter underneath the defect region. The post is required to form a conduction channel for charge carriers to travel from the substrate to the photonic crystal membrane. More detailed discussion of this photonic crystal structure will be presented in a later section.

It may also be difficult for suspended membrane structures to satisfy industrial standards for reliability. Burying the membrane in oxide or other low-index dielectrics would provide better mechanical support, help passivate the surface (suppressing surface recombination of free carriers, among other advantages), and improve thermal dissipation characteristics.

Certainly, process integration involves a range of issues well beyond the scope of this short section. However, there appears to be no fundamental barrier to achieving economical mass production of photonic crystal optoelectronic chips.

11.7 Photonic Crystal Light-Emitting Devices and Lasers

Spontaneous emission from an excited atom is not an intrinsic property of the atom itself; rather, it is the result of interaction of the atom with the radiation field surrounding it. As such, spontaneous emission can be controlled by modifying the radiation field. A common way is to enclose an atom in a limited

space, or cavity, to isolate it from the larger space outside. The radiation field inside the cavity occupies certain cavity modes, which exist only in particular frequency ranges. If the optical transition frequency of the excited atom does not coincide with the frequency of a cavity mode, emission is inhibited. If it does coincide with the frequency of a cavity mode, emission is allowed. Moreover, in 1946, Edward Purcell proposed that because the radiation mode density of a resonant cavity can be made significantly higher than that of vacuum, the spontaneous emission from an atom enclosed in a cavity can be significantly enhanced [73]. The enhancement factor, later named the Purcell factor, is

$$F_{\mathrm{p}} = \frac{3Q(\lambda/n)^3}{4\pi^2 V_{\mathrm{eff}}} \tag{11.27}$$

where Q is the quality factor of the cavity, λ is the free-space wavelength, n is the refractive index of the medium inside the cavity, and V_{eff} is the effective mode volume [74].

Photonic crystal cavities are capable of providing an ultrasmall effective mode volume V_{eff} of merely a few times $(\lambda/2n)^3$ [75] and can give a quality factor as high as $Q{\sim}10^6$ [30,31]. If these two numbers can be simultaneously achieved in a single device, an ultrahigh Purcell factor F_{p} ~500,000 can be obtained at the resonance frequency. Since the stimulated emission rate is proportional to the spontaneous emission rate, a high Purcell factor implies that the stimulated emission is also significantly enhanced. An enhancement in stimulated emission lowers the lasing threshold.

11.7.1 Optically Pumped Photonic Crystal Cavity Lasers

Early work on photonic crystal cavity lasers used optical pumping of the cavity mode [75]. The structure used was a suspended photonic crystal slab membrane with a hexagonal lattice of holes. For many low-threshold laser applications, it is desirable to have a single-mode cavity. Due to the symmetry of the lattice, a cavity made by filling a single hole has degenerate dipole modes. To break the symmetry and create a single-mode cavity, a pair of holes was enlarged instead. The estimated quality factor and mode volume were 250 and $2.5(\lambda/2n)^3 = 0.03\,\mu\mathrm{m}^3$, respectively. The photonic crystal was fabricated in an InGaAsP film grown by metal–organic chemical vapor deposition (MOCVD) on an InP substrate. The active region comprised four 9 nm InGaAsP quantum wells separated by 20 nm quaternary barriers with 1.22 μm band gap. The quantum well emission wavelength was designed for 1.55 μm at room temperature. The compressive strain in the InGaAsP quantum wells enhances coupling to the TE polarization, which favors the actual defect mode of the photonic crystal cavity. The hexagonal photonic crystal lattice structure was patterned by e-beam lithography and dry-etched using a metal/SiN double hard-mask layer. The

InP sacrificial layer directly beneath the photonic crystal was removed by dipping in a diluted HCl solution. A free-standing photonic crystal membrane was thus obtained.

Lasing action was demonstrated at a substrate temperature of 143 K for pulses of 10 ns wide and 250 ns apart. Above the threshold pump level of 6.75 mW, the line width shrunk from 7 nm to below 0.2 nm, the latter number being limited by the resolution of the spectrometer. However, because of the low quality factor and relatively large pumping beam (30 times greater area than the defect mode), the threshold of this first optically pumped laser was quite high.

Recent work has demonstrated very fast photonic crystal slab lasers with response times as short as a few picoseconds [76].

Laser cavities have also been fabricated in 3D woodpile photonic crystals by the wafer fusion technique [77]. InGaAsP multiple quantum wells were introduced as the active layer into GaAs photonic crystals with an in-plane period of 0.7 μm and a stripe width and height of 0.2 μm. Without artificial defects, photoluminescence was suppressed in the photonic band gap. The inclusion of defects introduced modes into the band gap around 1.55 μm. As the sizes of the defects were reduced, individual cavity modes became distinguishable through photoluminescence measurements. Quality factors above 100 were achieved in defects of $0.76 \times 0.65\ \mu m^2$.

11.7.2 Electrically Pumped Photonic Crystal Lasers

Electrical pumping is preferable to optical pumping for many laser applications. However, an electrically pumped photonic crystal laser requires an intricate design of the carrier injection region so as to maintain a high quality factor, a small mode volume, and, often, a single-mode cavity. While these requirements are not easy to realize for optically pumped photonic crystal lasers, they become even more challenging in the presence of semiconductor carrier injection structures.

Electroluminescence from a photonic crystal defect structure was observed in 2000 [78,79]. The active layer of the device was a pair of 7 nm $In_{0.15}Ga_{0.85}As$ quantum wells surrounded by $Al_{0.3}Ga_{0.7}As$ layers. Additional p-type and n-type $Al_{0.96}Ga_{0.04}As$ layers were inserted above and below the quantum well layers for lateral wet oxidation, similar to the structure of an oxide-confined vertical cavity surface-emitting laser (VCSEL). The device has a bottom distributed Bragg reflector (DBR), but the top DBR is absent. The triangular photonic crystal has a lattice constant of $a = 0.4$ μm and an air hole radius of 0.13 μm. Deep dry etching was employed to produce 0.8 μm deep holes, such that the quantum well layers were roughly located at half the depth of the air holes. Wet oxidation of the periphery of the $Al_{0.96}Ga_{0.04}As$ layer cuts off the vertical carrier transport in the periphery, efficiently funneling carriers

horizontally toward the photonic crystal cavity. This results in a better spatial overlap of the gain region and the cavity mode profile, suppressing the potential spontaneous emission from the active layer in the photonic crystal region surrounding the center cavity. The opening inside the oxide ring is approximately 40 μm in diameter. The photonic crystal occupies most of this area.

Because the depth of the air holes is substantially larger than the cavity height, the cavity tends to be multimode. Electroluminescence measurements with 1 μs pulses at 1% duty cycle showed that the emission spectra remained broad above the threshold. Due to the absence of a top DBR mirror, the cavity modes had a low vertical $Q{\sim}12$. As a result, line widths of the cavity modes overlapped in frequency even above threshold. Near-field images showed that the mode intensity is confined to a 4 μm area, which excludes the possibility that the large 2D photonic crystal area surrounding the cavity contributes to the observed luminescence. However, the current at this "soft threshold" was relatively low (300 μA). In contrast to the membrane photonic crystal laser, this structure has excellent heat conduction through the substrate, allowing for room temperature operation for much longer pulses. The far-field radiation pattern had a full width at half maximum (FWHM) of 30°, smaller than the value of 90° observed for larger oxide-confined light-emitting diodes. Polarization measurements indicated a preferential direction.

A single-mode electrically pumped photonic crystal laser was developed by a Park et al. [72]. Through intricate design, this device simultaneously achieved small mode volume, high quality factor, single-mode operation, and particularly high spatial overlap of gain and optical mode profile. The design includes a carrier-transporting central post located just below the cavity defect region. This submicron-diameter post serves as "an electrical wire (for holes), a mode selector, and a heat sinker at the same time" [72]. One key advantage of the post is to limit the lateral diffusion of one type of carriers (in this case, holes). Electrons must be transported to the center defect region to recombine with holes. This effectively suppresses the spontaneous emission from the large photonic crystal area surrounding the cavity and significantly improves the spontaneous emission factor β (the portion of the spontaneous emission in the desired mode) of the device. In this sense, this post functions as a mode selector that favors spontaneous emission into a non-degenerate monopole cavity mode. Moreover, because the monopole mode has an intensity minimum at the center of the cavity, the introduction of the central post has minimal impact on the mode [80].

It is challenging to control the position, size, and shape of the post in fabrication because no lithography technique can reach below the surface to directly pattern the post. Instead, the geometric parameters of the post are controlled by manipulating the size of the air holes near the central defect, which in turn controls the wet etching speed that is critical to shaping the post. The manipulation of air holes results in five heterogeneous photonic crystal lattices with the same lattice constant but different air hole sizes. This chirped

photonic crystal cavity retains a monopole mode with an unchanged mode volume and frequency and a slightly improved Q.

The fabricated cavities were simulated using structural parameters obtained from an SEM image of the actual device. The theoretical value of $Q{\sim}3480$ agrees reasonably with measured values of $Q \sim>2500$. The theoretical value of the mode volume $V_{\text{eff}} = 0.684(\lambda/n)^3$ approaches the theoretical lower limit. This leads to an estimated Purcell factor ~270. The single-mode photonic crystal laser has a low threshold current of 260 μA and a turn-on voltage less than 1.0 V. A soft turn-on shoulder near the threshold gives a large spontaneous emission factor $\beta{\sim}0.25$ by fitting the L–I curve against the theory. However, owing to non-radiative recombination at the air–semiconductor interfaces in the air holes and at the edge of mesas, the external quantum efficiency of the laser is yet to be improved.

11.7.3 Band-Edge Lasers: In Planar Structures and in Waveguides

The low group velocity near a photonic band edge can significantly enhance the gain in a photonic crystal structure. An expression for the gain coefficient can be derived by studying the response of a photonic crystal mode to perturbation by optical emission from impurity atoms [81]. Sakoda's approach is to study the optical response to a perturbation in the form of optical emission from impurity atoms into a photonic crystal mode, $\mathbf{E}_{m\mathbf{k}}(\mathbf{x})$. This perturbation can be conveniently described by a polarization vector

$$\mathbf{P}_{st}^{(1)}(\mathbf{x}, t) = \alpha N(\mathbf{x})\mathbf{E}_{m\mathbf{k}}(\mathbf{x}) \exp(-\mathrm{i}\omega t + \delta t) \tag{11.28}$$

where $N(\mathbf{x})$ is the density of impurity atoms, α is their polarizability, and δ is a small positive number introduced to ensure convergence. The displacement vector in the presence of $\mathbf{P}(\mathbf{x},t)$ can be calculated using the Green's function, which gives

$$\mathbf{D}(\mathbf{x}, t) = \frac{\varepsilon(\mathbf{x})}{V} \sum_{n\mathbf{k}} \mathbf{E}_{n\mathbf{k}}(\mathbf{x}) \int \mathrm{d}\mathbf{x}' \int \mathrm{d}t' \mathbf{E}_{n\mathbf{k}}^*(\mathbf{x}')\mathbf{P}(\mathbf{x}', t) \times \omega_{n\mathbf{k}} \sin[\omega_{n\mathbf{k}}(t - t')] \tag{11.29}$$

where $\omega_{n\mathbf{k}}$ is the frequency of mode $\mathbf{E}_{n\mathbf{k}}$ and V is the volume of the photonic crystal. On the other hand, we can write $\mathbf{D}(\mathbf{x},t) = \epsilon_0\epsilon(\mathbf{x})\mathbf{E}(\mathbf{x},t) + \mathbf{P}(\mathbf{x},t)$. Thus, a self-consistent equation can be solved to obtain

$$\mathbf{E}(\mathbf{x}, t) = \mathbf{E}_{m\mathbf{k}}(\mathbf{x}) \exp(\tilde{g}_{m\mathbf{k}} L_z - \mathrm{i}\omega_{m\mathbf{k}} t)$$

and the gain coefficient (for optical power) is given by

$$g_{m\mathbf{k}} = 2\mathrm{Re}[\tilde{g}_{m\mathbf{k}}] = -\frac{\mathrm{Im}[\alpha]\omega_{m\mathbf{k}} N_{\mathrm{eff}}}{\varepsilon_0 v_{\mathrm{g}}} \tag{11.30}$$

where $v_{\mathrm{g}} = \partial\omega_{n\mathbf{k}}/\partial k_z$ is the group velocity and N_{eff} is the effective impurity density. The gain coefficient, $g_{m\mathbf{k}}$, increases significantly as the group velocity approaches zero.

In the above analysis, it is assumed that there is one preferred mode $\mathbf{E}_{m\mathbf{k}}$ whose group velocity is along the z-axis. If there is no preferred mode (e.g., for spontaneous emission), then the summation in Eq. (11.29) must include all photonic crystal modes at frequency ω. The integrand with respect to the frequency will have the density of states $D(\omega)$ as a factor. Depending on the dimensionality, a vanishing group velocity may not always enhance spontaneous emission [82]. This phenomenon is related to the nature of Van Hove singularities in various dimensions. A well-known example is that in a 2D lattice, the Van Hove singularity may take the form of a finite jump in the density of states (for band maximum or minimum) rather than a divergent $D(\omega)$. Spontaneous emission will not be enhanced in such a case. Detailed experimental and theoretical investigation was conducted to illustrate this effect [82]. Interestingly, for a system with 1D periodicity [83], the band-edge enhancement is, however, present in general.

11.7.4 Application to the Extraction Efficiency of Light-Emitting Diodes and VCSELs

Another interesting application of photonic crystals is light-emitting diodes (LEDs). Conventional semiconductor LEDs suffer from poor light extraction efficiencies due to total internal reflection at the semiconductor–air interface. For a semiconductor refractive index of $n = 3.4$ (corresponding to a half-cone angle of $\sin^{-1}(1/n) = 17°$), the efficiency is $1/4n^2 \cong 2\%$ [84]. Placing the active layer of the LED in a 2D photonic crystal slab can enhance the extraction efficiency [85]. The photonic crystal should ideally be designed so that the LED emission spectrum falls in a frequency range for which all modes of the photonic crystal slab radiate to air. For a PC slab of sufficiently large area, close to 100% efficiency is expected theoretically [85].

Experiments have observed a sixfold enhancement of photoluminescence in a InGaAs/InP double heterostructure [84]. However, the penetration of air holes into the active layer of a LED causes additional surface recombination, lowering the internal quantum efficiency. An alternative scheme places a 2D photonic crystal grating *above* the active layer of a LED [86]. If the photonic crystal layer is sufficiently shallow, the enhancement of the extraction efficiency can be described by the grating diffraction effect [87].

Photonic crystal structures have also been used in vertical cavity surface-emitting lasers (VCSELs) [88]. A photonic crystal was incorporated into the top

distributed Bragg reflector (DBR), and an oxide aperture in the bottom DBR was used to restrict lateral current spreading beyond the area covered by the photonic crystal defect. Single-mode operation was demonstrated with sub-milliamp threshold current and a milliwatt of output power.

11.8 Photonic Crystal Filters

By combining cavities and waveguides in photonic crystals, it is possible to create different types of optical filters [89]. Such devices may find application in optical communications, particularly in wavelength division multiplexing (WDM). In WDM systems, signals are encoded on multiple channels, each of which occupies a separate frequency bandwidth. Optical filters that can separate out and redirect particular channels from the optical data stream are useful for optical processing of the signal [90]. In comparison to alternate technologies such as micro-ring resonators, photonic crystal devices are extremely compact in size, allowing denser on-chip integration.

One basic filter design is shown schematically in Fig. 11.17. It consists of a waveguide side-coupled to a microcavity resonator. For simplicity, we represent the waveguide with a thick solid line and the microcavity resonator by a solid ellipse. The actual structures used to implement waveguide and microresonator elements could resemble, for example, the photonic crystal slab waveguides and microcavities discussed earlier in the chapter.

Transmission through the system can be derived using coupled-mode theory, a general and flexible framework for describing the transfer of light between electromagnetic modes in time [91]. We assume that the microcavity resonator supports a standing wave at frequency ω_o. The radiative loss of the resonator to the outside world is described by a radiative quality factor Q_{rad}, and the coupling between the resonator and waveguide can be described by a coupling

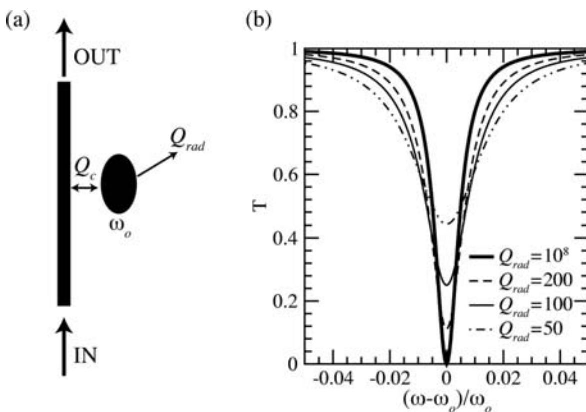


Fig. 11.17 (**a**) Schematic diagram of band-rejection filter. (**b**) Transmission

quality factor Q_c, which is inversely proportional to the resonator–waveguide coupling rate. The transmission is calculated to be [92]

$$T = \frac{P_{out}}{P_{in}} = \frac{(\omega - \omega_o)^2 + \left(\frac{\omega_o}{2Q_{rad}}\right)^2}{(\omega - \omega_o)^2 + \left(\frac{\omega_o}{2Q_{rad}} + \frac{\omega_o}{2Q_c}\right)^2}$$

The transmission spectrum for varying values of Q_{rad} is plotted in Fig. 11.17(b). The transmission spectrum has the form of a dip, with the minimum at the resonant frequency of the microcavity. The structure functions as a band-rejection filter. The linewidth of the filter decreases with increasing Q_c. For infinite Q_{rad}, the transmission is zero at ω_o. As Q_{rad} decreases, the minimum transmission increases, degrading the filter performance. The light that is not transmitted is partially reflected and partially emitted into radiation modes.

Since the emitted light lies within the rejection bandwidth of the filter, the device can be said to "drop" light in this frequency range. In photonic crystal slab implementations known as *surface-emitting channel drop filters*, the dropped light radiates above and below the slab, allowing it to be collected [93]. By cascading a sequence of filters centered at different resonant frequencies ω_i, different frequency ranges of the signal can be spatially separated, as shown schematically in Fig. 11.18. For each individual filter, the line width of the drop spectrum can be adjusted by changing the separation distance between the waveguide and cavity; larger separations correspond to weaker coupling (higher Q_c) and narrower line widths [94,95]. The minimum resolution of the filter is limited by Q_{rad}, the radiative quality factor of the cavity. Noda and co-workers have experimentally demonstrated optimized structures with resolution as high as 0.25 nm [95], as well as multichannel dropping of channels with resolution of 0.4 nm [96]. They have further demonstrated that the drop wavelength of a filter can be dynamically tuned using thermal heating [97].

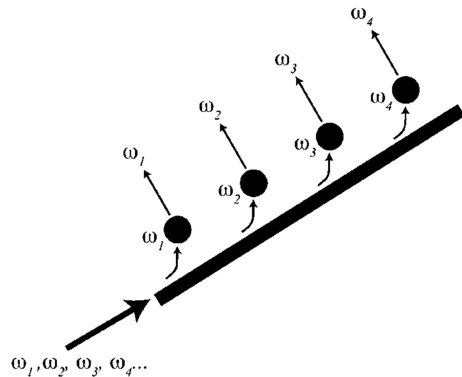


Fig. 11.18 Schematic operation of a surface-emitting channel drop filter for separating multiple frequencies

One drawback of surface-emitting devices is the practical complication associated with collecting vertically emitted light. An alternative class of designs is the in-plane channel drop filters.

A simple design for an in-plane channel drop filter is shown in Fig. 11.19(a). A microcavity resonator is placed between two waveguides. Light entering the input waveguide can tunnel into the resonator mode and exit through the drop port. However, the efficiency of this process is only 25%. Fan et al. showed that it is possible to increase the efficiency by using a double-resonator system, as shown in Fig. 11.19(b) [89,98]. The structure is designed to support two resonator modes at the same frequency, one that has even symmetry with respect to the center plane of the structure and one that has odd symmetry. Light in the input waveguide can excite both modes. Due to interference between the opposite-symmetry degenerate modes, it turns out that 100% of the input light is transmitted to the drop waveguide, assuming an ideal, lossless system ($Q_{\text{rad}} \to \infty$).

Subsequent work has presented detailed theoretical designs for channel drop filters in photonic crystal slabs [99,100]. One major challenge for filter performance is the necessity of obtaining high values of Q_{rad} relative to Q_{c}. After careful optimization of the photonic crystal structure, theoretical channel selectivity in the 0.2 nm range can be obtained. However, one practical difficulty with filter designs based on doubly degenerate modes is that even very slight imperfections in the fabricated photonic crystal structure will disturb the degeneracy, causing the two modes to split in frequency and compromising filter performance. Later designs used the interference between reflection from a photonic crystal heterostructure interface and reflection from a single cavity to achieve a theoretical efficiency of 100% in the absence of radiation loss [101,102]. Four-channel drop operation has been demonstrated experimentally with high efficiencies of almost 100% [103]. Alternately, drop filters based on tunneling from input to output through a single cavity have also been introduced [104].

Waveguide–resonator systems of higher rotational symmetry were also investigated [105]. Multichannel add-drop filters were designed and their performance was calculated. A new mechanism was presented to reduce the remnant light at the dropped wavelengths in the pass-through port. Generally,

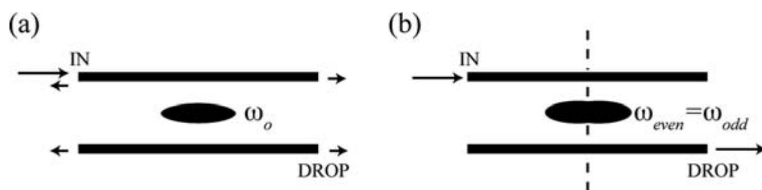


Fig. 11.19 In-plane channel drop filters: (**a**) simple, suboptimal design, (**b**) design with 100% theoretical transmission efficiency

in a WDM system, when a wavelength is dropped from an optical fiber through an add-drop filter, it is desired that the drop process leave behind a wavelength channel as clean (or empty) as possible.

Therefore, the reduction of remnant light through this approach is of great interest for WDM applications. High-order Butterworth filters that have better flat-top filter profiles can be also achieved in these high-symmetry systems. Such flat-top filters suppress crosstalk (i.e., inter-channel interference/noise) between adjacent wavelength channels and enlarge the effective bandwidth of each wavelength channel. These merits are also highly desirable for WDM applications.

11.9 Photonic Crystal Modulators

In this section, we discuss photonic crystal Mach–Zehnder modulators. Photonic crystal waveguides (PCWs) can slow down the speed of light by up to 1000 times [106,107]. Slow light speeds increase the change in phase velocity for fixed propagation length. The phase modulation efficiency is thus significantly enhanced and the modulator electrode length is reduced by several orders of magnitude [108]. Recent experimental demonstrations [109,110,111] have shown great promise of utilizing such modulators in silicon-based optical interconnects. Optical interconnects are of interest for overcoming the electronic interconnect bottleneck faced by the microelectronics industry.

As we see from the case of photonic crystal modulators, while electrical engineering was an offspring of physics, it continues to benefit from physics for inspiration. The development of photonic crystal modulators illustrates the engineering challenges that arise when a new scientific idea moves into a practical device realization. Such work calls for scientists and engineers to look outside their specific domains of expertise and work across disciplinary boundaries to foster innovation.

11.9.1 Basic Idea: Slow Light Enhancement of Phase Shift

A schematic of a photonic crystal waveguide modulator is depicted in Fig. 11.20. As in a standard Mach–Zehnder modulator, light is split between two waveguide arms, one of which has an electrically tunable phase shift. The two light beams recombine constructively or destructively depending on the phase difference between the arms. Modulating the phase difference modulates the intensity of the output signal. In the figure, the photonic crystal waveguide is formed on the top silicon layer of a silicon-on-insulator wafer.

We consider a typical dispersion relation of a photonic crystal waveguide mode, shown in Fig. 11.21(a) [110]. When the refractive index of the core material changes slightly, the dispersion curve shifts vertically by an amount

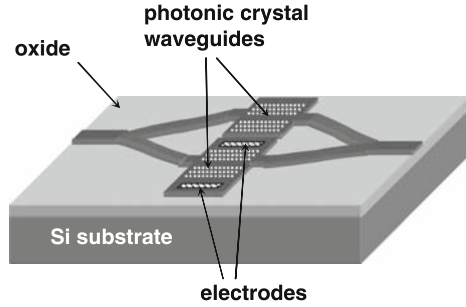


Fig. 11.20 Schematic of a photonic crystal modulator on a silicon-on-insulator substrate

$\Delta\omega_0 \sim \omega(\Delta n/n)$. Consider the effect for a fixed frequency (or, equivalently, a fixed wavelength). The change in the propagation constant is related to the frequency shift through a factor inversely proportional to the group velocity,

$$\Delta\beta_{\mathrm{PC}} = \Delta\omega_0 / v_{\mathrm{g}} \tag{11.31}$$

The phase shift across a segment of PCW of length L can be expressed as $\Delta\phi = \Delta\beta_{\mathrm{PC}} \cdot L$. Therefore, the interaction length required to obtain a π phase shift for a guided mode is

$$L \sim \frac{n}{2\Delta n} \cdot \frac{v_{\mathrm{g}}}{c} \lambda_{\mathrm{air}} \tag{11.32}$$

where λ_{air} is the wavelength in air. It is evident that when the group velocity approaches zero near the band edge, the interaction length required to achieve a given phase shift can be reduced significantly, as first proposed by Soljacic et al. [108].

The difference between a photonic crystal waveguide and a homogeneous medium is illustrated in Fig. 11.21. The dispersion curve of a homogeneous medium is shown in Fig. 11.21(b). A change of refractive index causes the linear

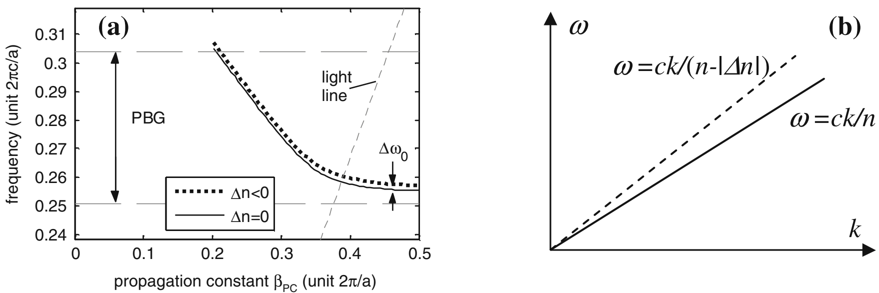


Fig. 11.21 Change of dispersion relation in response to an index perturbation. (**a**) Dispersion relation of a photonic crystal waveguide; (**b**) dispersion relation of a homogeneous medium

dispersion relation to change its slope, pivoting around the origin of the ω–k diagram. The change of wavevector (or propagation constant) is simply $\Delta k = k_0 \Delta n$. For a photonic crystal waveguide (Fig. 11.21(a)), the periodicity of the dispersion relation $\omega(\beta_{PC} + 2\pi/a) = \omega(\beta_{PC})$ along with inversion/mirror symmetry along the waveguide axis ensures an extremum (maximum or minimum) of $\omega(\beta_{PC})$ at $\beta_{PC} = \pi/a$, where $v_g = \omega(\beta_{PC}) = 0$. A perturbation of the refractive index can shift the curve vertically or change its curvature ("effective mass"), but the extremum must remain at $\beta_{PC} = \pi/a$, assuming the lattice constant a does not change. The change in curvature is usually negligible for a small perturbation of the refractive index. Therefore, the change of β_{PC} is mainly due to the vertical shift of the dispersion curve and can be very large for frequencies in the flat (low group velocity) portion of the band, near the band edge.

An intuitive interpretation is that light travels slower in a PCW than in a bulk material and has more time to interact with electrons. This enhances light–matter interaction and shrinks the interaction length.

The original proposal for slow light photonic crystal waveguide modulators [108] was to use coupled-cavity photonic crystal waveguides (CCWs). However, CCWs in photonic crystal slabs have high intrinsic optical loss, making line-defect waveguides experimentally preferable. CCWs tend to be intrinsically lossy because the longitudinal period is n ($n \geq 2$) times the original lattice constant, which makes the Brillouin zone along the waveguide axis n times smaller. For the 2D photonic crystal slab configuration, the remnant gap of the photonic crystal generally lies above the light line at the BZ boundary $k = \pi/(na)$ of the CCW, while it lies below the light line at the BZ boundary $k = \pi/a$ of an ordinary photonic crystal waveguide.

11.9.2 Passive Photonic Crystal Waveguide Mach–Zehnder Interferometers

Passive Mach–Zehnder interferometers (MZI) formed in 2D photonic crystal structures were reported in 2004 [71]. The structures were fabricated on an InGaAsP layer deposited on an InP substrate. The photonic crystal structures were patterned by e-beam lithography, followed by electron cyclotron resonance etching using an Au/Cr layer as a hard mask. A suspended membrane was formed by HCl/H_2O wet etching. To assess the effect of the slow group velocity on the phase shift, Mach–Zehnder modulators with both asymmetric and symmetric arms were fabricated. The experimental transmission data as a function of the inverse wavelength, $T(1/\lambda)$, were Fourier transformed to accurately identify the periodicity of the spectra. The phase change for an asymmetric MZI whose two arms have a length difference of ΔL is given by

$$\Delta\phi = \omega_0 \Delta L / v_g = (2\pi/\lambda) \Delta L \cdot n_g \quad (11.33)$$

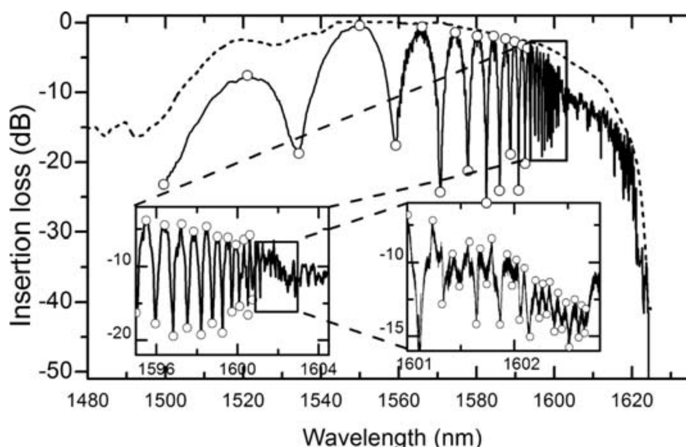


Fig. 11.22 The transmission spectra of a photonic crystal waveguide (*dashed*) and a photonic crystal Mach–Zehnder interferometer (*line*) with two slightly asymmetric arms. *Insets* show the fine details of the MZI spectrum (courtesy of Yurii A. Vlasov)

Since the intensity spectrum is given by $T(1/\lambda) = \cos(\Delta\phi)$, the Fourier transform of the spectrum will have a peak at $n_g \Delta L$. Reflection from the waveguide end facets or other structural features can impose additional oscillation periods on the spectrum. The Fourier transform allows the different oscillation periods to be separated [71].

Accurate characterization of the *wavelength-dependent* slow group velocity near the band edge of a guided mode mandates a delicate experimental effort, which was completed at IBM T.J. Watson Research Center [109]. With a carefully fabricated asymmetrical MZI, Vlasov et al. observed, with high accuracy, the shortening of fringe oscillation periods as the transmission approaches the band edge (Fig. 11.22). Their experiments unequivocally established the enhancement of phase shift sensitivities to the group velocity in a photonic crystal waveguide.

11.9.3 Engineering a Photonic Crystal Waveguide Modulator

While the scientific principles of a photonic crystal waveguide modulator have been known since 2002, the design and fabrication proved challenging. For a device to be useful in communications, high-speed modulation well beyond a kilohertz is desired. A combination of physics and engineering knowledge in photonics, electronics, and heat transfer is required to design a photonic crystal modulator that outperforms conventional modulators. Indeed, among a number of photonic crystal waveguide modulators fabricated so far, only a couple of them have demonstrated evident improvement over conventional modulators [109,110,111].

The impetus for photonic crystal modulator research has been fueled by the burgeoning interest in silicon-based photonics, especially high-speed silicon modulators [112]. The wide availability of relatively inexpensive silicon-on-insulator wafers and the enormous potential of optoelectronic integration on silicon have made silicon a favorable material for modulator research. We will focus on silicon modulators in the subsequent discussion, although many design and fabrication considerations can be applied to other semiconductor modulators as well.

The most important performance indices for an optical modulator are modulation depth, optical bandwidth, insertion loss, half-wave voltage V_π (usually measured at DC), electrical bandwidth, and driving current or power consumption. In addition, if optical modulators are to be integrated into planar lightwave circuits that can be mass manufactured with today's VLSI technology, the modulator design must be compatible with the prevailing processing and packaging technology, the details of which are beyond the scope of this chapter.

Although it is possible to fabricate a Mach–Zehnder interferometer made entirely of photonic crystal waveguides, such a structure is complicated for a number of reasons. The key advantage of introducing photonic crystal waveguides into a Mach–Zehnder modulator is the reduced interaction length, i.e., the length of the waveguide segment that is subject to electrical tuning of the refractive index. It is the interaction length, not the overall length of the MZI, that relates to critical issues such as voltage and power consumption of the modulator. For initial demonstrations, it is reasonable to use a photonic crystal waveguide only for this electrically controlled segment so as to reduce the design complexity of the device. This idea gained popularity in many early demonstrations [109,110,111,113].

A number of design issues arise from the electrical structure of a modulator. There are many schemes for injecting electrons and holes into a photonic crystal waveguide. Generally, the electrical structures can be divided into vertical configurations or horizontal configurations according to their geometry and MOS capacitors or p-i-n diodes according to their electrical implementation. Early work on silicon modulators favored the p-i-n diode configuration, whereas the Intel group [112] advanced the MOS capacitor structure.

A vertical configuration is shown in Fig. 11.23(a). The key design problem stems from the top electrode. To reduce the optical loss, a thick poly-Si layer must be inserted underneath the top electrode to ensure that the tail of the optical mode field does not strongly overlap with the electrode. The thick poly-Si layer may cause a W1 waveguide to have multiple modes, which is undesirable for a MZI. It is in principle possible to reduce the waveguide width and enforce the single-mode condition. However, this generally results in a poly-Si structure with high aspect ratio, which may cause difficulties in planarization or other processing steps. In addition, the electric current flows along the longitudinal direction of the electrode, entering and exiting through the two ends of the metal wire. The effective cross-section for the current flow equals the

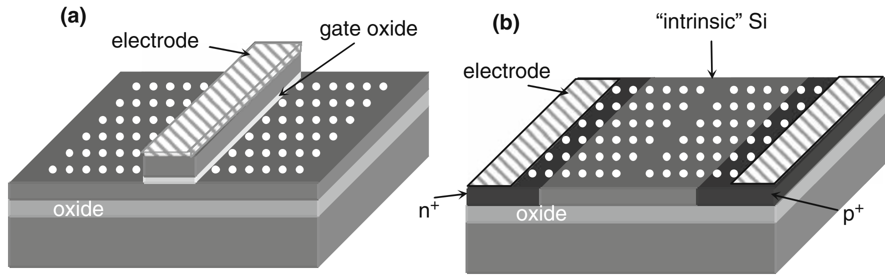


Fig. 11.23 Two electrical configurations of a silicon photonic crystal waveguide modulator. (**a**) A vertical MOS capacitor embedded in a photonic crystal waveguide; (**b**) a horizontal p-i-n diode embedded in a photonic crystal waveguide

waveguide width times the metal thickness. Issues such as electromigration limit the current density. Since the waveguide width is on the submicron scale, the cross-section is too small to accommodate a high current, limiting the modulation depth. Further study is needed to understand whether these problems can be overcome within the vertical configuration.

A horizontal configuration, depicted in Fig. 11.23(b), avoids most of the problems associated with the vertical configuration and allows us to take advantage of the extensive existing knowledge on standard photonic crystal waveguides. In addition, a horizontal p-i-n diode is more planar than a vertical MOS capacitor, shown in Fig. 11.23(a). The planarization advantage is of critical importance for fabrication and integration of a modulator with microelectronic circuits for optical interconnects and other on-chip applications. The MOS capacitor configuration usually gives a thin layer of charge carriers that overlap with a very small portion of the optical mode field. This is not conducive to enhancing the interaction between light and electrons. Therefore, we consider the p-i-n diode as the first choice.

P-i-n diode-based modulators are considered to be slower than MOS-based modulators [112]. In most silicon modulators, the carrier generation process has a negligible effect on high-speed modulation. The key carrier transport/transition processes affecting high-speed modulation include carrier recombination, diffusion, and drift. For moderate to high forward injection levels, the diffusion process provides the main portion of the excess carriers and electric current. Upon a sudden switch to reverse bias in a modulation cycle, the junction voltage and internal field remain at relatively small values. Diffusion and recombination are important to expedite the removal of excess carriers in the i-region. Recent simulations and experiments have revealed that the removal of carriers under reverse bias is rapid for compact p-i-n diode modulators whose waveguide cross-sections are less than 0.5 μm × 0.5 μm, and the slow rising time under forward bias is the primary concern [114,115,116].

There is an important, yet frequently overlooked, *engineering* advantage of using photonic crystal waveguides in modulators. For a modulator in the

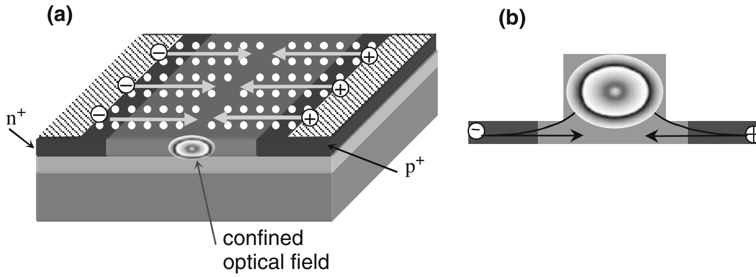


Fig. 11.24 Comparison of a photonic crystal waveguide modulator and a conventional silicon modulator, both in the horizontal configuration

horizontal carrier injection scheme, it is desirable for the waveguide cladding to be optically insulating but electrically conducting. In other words, light cannot leak through the cladding, but electrons/holes can be injected through it. A photonic crystal waveguide fulfills this requirement exactly, as shown in Fig. 11.24(a). In contrast, in a conventional modulator, it is necessary to etch the side of a rib waveguide to achieve optical confinement, or insulation, as shown in Fig. 11.24(b). This simultaneously reduces the electrical conductivity significantly. It also requires that electrons/holes diffuse vertically to overlap with the peak of the optical field. These factors can significantly compromise the modulation efficiency and/or speed.

11.9.4 High-Speed, Low-Voltage Silicon Photonic Crystal Modulator

A key issue for high-speed silicon modulators is the driving voltage. Gigahertz silicon modulators typically have a peak driving voltage well above 5 V [112,116]. Such a high driving voltage and the accompanying high power consumption are undesirable for most on-chip applications. It turns out that the high voltage can be attributed to certain intrinsic properties of silicon and the fundamental limit on the size of a guided optical mode.

Intensity modulators made of silicon generally need to dynamically produce a critical carrier concentration perturbation on the order of $(\Delta N_e)_c = (\Delta N_h)_c = 3\times10^{17}\,\mathrm{cm}^{-3}$ [117,118,119]. With this general requirement, we examine the speed scaling of silicon p-i-n diode modulators. Consider an arbitrary optical waveguide whose core width is on the order of $w_{\mathrm{core}} = 1\,\mu\mathrm{m}$ for wavelengths around 1.55 μm. For moderate to high forward injection, we assume that *everywhere* in the diode, the excess carriers are non-decreasing during the forward-bias stage. Also, we assume that the carrier generation is negligible for moderate to high injection levels. Therefore, the excess carriers are ultimately supplied externally by the injected current. Regardless of the detailed carrier

transport mechanism, the time required to fill the waveguide core to an optically critical level of $(\Delta N_h)_c$ cannot be shorter than

$$\Delta t = q w_{\text{core}} (\Delta N_h)_c / J \tag{11.34}$$

where J is the current density and q is the electron charge. Note that a similar formula has been used to analyze a case where the recombination process prevailed [118]. This limit reduces to the well-known transit time limit if the drift current dominates: $J \sim q(\Delta N) v_d$, where v_d is the drift velocity.

The importance of Eq. (11.34) lies in the fact that the quantities q, w_{core}, and $(\Delta N_h)_c$ are either fixed or have certain limits set by fundamental physical laws. For example, the lower limit of the waveguide core width (or more accurately, the mode field width) is on the order of the wavelength due to the fundamental nature of light. The optically critical carrier concentration is an intrinsic property of silicon. For these reasons, the carrier density limit can be considered a fundamental limit for a wide range of silicon-based intensity modulators, including the MZI and directional coupler [120].

The scaling of Eq. (11.34) may also be expressed in terms of the modulation frequency f. Assume that the filling time is half a period, $\Delta t = 1/(2f)$. Then

$$J = 2 q w_{\text{core}} (\Delta N_h)_c f \tag{11.35}$$

A simple calculation shows that Eq. (11.35) requires a current density on the order of 10^4 A/cm^2 for $f = 1$ GHz. For a conventional waveguide modulator, the cross-section for the electrical current is $A = hL$. Assuming a waveguide height h above 1 μm and a waveguide length L around 1 mm, the required current is above 0.1 A. Note that a vertical diode setting will reverse h and w_{core}, but the conclusion remains the same. Even if a conventional silicon modulator can achieve a low impedance value of 50 Ω, the required power and voltage may not be acceptable for most on-chip optical interconnect applications [121,122].

Scaling down the device dimensions can be the answer to this difficulty. Photonic crystal-based structures can shrink the device interaction length to tens of microns and the device height to hundreds of nanometers. This significantly reduces the overall current for the same current density. In contrast, in the vertical MOS capacitor configuration mentioned above, the cross-section of the electric current is $A \sim w_{\text{core}} h$. Since $w_{\text{core}} << L$, the overall current for a given maximum current density is significantly limited, an undesirable feature for high-speed modulators.

The injection level $J \sim 10^4$ A/cm^2 required for gigahertz modulation falls in the high injection regime of a diode. This causes "slower" carrier concentration increase with the junction voltage V_j in the form of

$$\Delta N_e = \Delta N_h = n_i \exp(q V_j / 2 k_B T)$$

As V_j approaches the contact potential V_0, this gives $\Delta N_e = \Delta N_h = (N_a N_{di})^{1/2}$, much lower than $\Delta N_h = N_a$ for an ideal diode. Here N_a is the acceptor concentration of the p-region and N_{di} is the donor concentration of the *i*-region. Usually, we have $N_a > 10^{19}$ cm^{-3} and N_{di}: 10^{15}–10^{17} cm^{-3}. Therefore, $(N_a N_{di})^{1/2} << N_a$.

Fabrication of the high-speed silicon photonic crystal waveguide modulator shown in Fig. 11.20 was completed with common microelectronics processing techniques. The optical waveguide layer (both the conventional silicon waveguides and the photonic crystal waveguides) was patterned by electron-beam lithography and dry etching. A critical problem for photonic crystal devices that involve electron transport is the surface recombination on the etched side walls of the air holes. In most cases, this recombination is undesirable for device performance and must be suppressed. For silicon-based photonic crystal devices, passivation by thermally grown oxide is the most straightforward approach. A typical passivation layer of 5–10 nm will cause a slight change in the dielectric structure and hence the guided mode band (such as a slight shift of the band edge). Generally, this effect must be considered along with other common fabrication tolerances such as the deviation of air holes.

A thin thermal oxide layer was grown to passivate the silicon surface. P^+ and N^+ regions were defined using photolithography and implanted to a concentration of about $N_a = N_d = 5\times10^{19}$ cm^{-3}. The intrinsic region was n-doped to $N_{di} \sim 10^{15}$ cm^{-3}. The aluminum electrodes were patterned by photolithography. To sustain high current density, care was taken to design the geometry of the highly doped regions and electrodes.

The optical characterization of the high-speed p-i-n diode-based MZI modulator was conducted for the transverse electric (TE) polarization at a wavelength of 1541 nm [111]. Polarization-maintaining (PM) fibers with lensed taper ends were used to couple light into the waveguides. A maximum modulation depth of 93% was obtained at a static injection current of 7.1 mA, indicative of low optical absorption under an injection level around 4×10^4 A/cm^2. Square wave input electrical signals having a peak-to-peak amplitude of 3 V (V_{on} = 2 V, V_{off} = −1 V) and a duty cycle of 50% were applied to the electrodes of the modulator. Thermo-optic modulation at these high frequencies is estimated to be insignificant compared to electro-optic modulation. A high modulation depth of 85% at 2 Mbit/s was obtained. The modulation depth was reduced by 3 dB as the modulation frequency increased to 1 Gbit/s, which marks the 3 dB bandwidth of our device.

11.9.5 Thermo-optic Photonic Crystal Waveguide Modulators

Thermo-optic Mach–Zehnder modulators have been demonstrated by several research groups [113,123,124]. Some devices employed a vertical configuration as shown in Fig. 11.25(a), where a heating electrode is placed above a dielectric

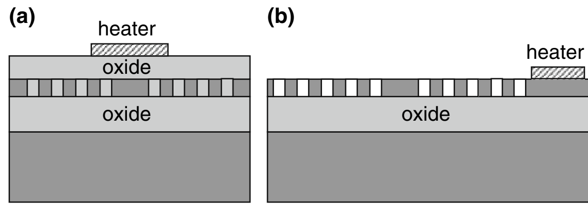


Fig. 11.25 Comparison of (**a**) vertical and (**b**) horizontal heater configurations

layer (e.g., silicon oxide) such that the heat flux is vertical. The dielectric layer must be considerably thick to separate the light-absorbing metal electrode from the waveguiding layer. However, a thick dielectric layer reduces the efficiency of heat transfer from the electrode to the waveguiding layer, because most dielectrics are poor conductors of heat.

The problem can be solved by adopting a horizontal (or in-plane) heating structure, shown in Fig. 11.25(b). Here the heating electrode is placed directly on the semiconductor layer in which the photonic crystal waveguide is formed. The thermal conductivity of semiconductors is generally higher than dielectric materials. For example, the conductivity of silicon is about two orders of magnitude higher than that of silicon oxide.

However, the semiconductor slab is usually very thin (thickness $t{\sim}0.6a$, where a is the lattice constant). In addition, it is found empirically that the electrode(s) must be separated from the waveguide by at least five rows of air holes to avoid high optical loss. As a result, although silicon has a high thermal conductivity, the improvement on the "thermal resistance," $\kappa d/tL$, in the horizontal configuration is not large, where d is the horizontal distance from heater to the waveguide core and L the waveguide length. However, an improvement factor of 3–30 is still possible relative to the vertical configuration. Switching time less than 1 μs has been achieved for a silicon-based thermo-optic modulator based on a photonic crystal waveguide Mach–Zehnder interferometer, and the switching power can be 2 mW or less [109]. These structures are promising for optical switching and optical storage applications. Further work is needed for the dynamic heat transfer process in silicon modulators.

11.10 Superprism Devices for Wavelength Demultiplexing and Sensing

Refraction from a photonic crystal surface is highly sensitive to wavelength and incident angle, a phenomenon known as the superprism effect [125]. In particular, the angular dispersion capability of a photonic crystal can be 500 times larger than a conventional prism. Photonic crystal superprisms that could separate a light beam of mixed colors into a large number of closely spaced wavelengths hold great interest for potential applications in high-bandwidth fiber-optic communications. In addition, the high sensitivity to wavelength in a

superprism is frequently accompanied by high sensitivity to refractive index change. This leads to promising applications in sensing and nonlinear optics. In this section, we will discuss the physical origin of the superprism effect and discuss several applications.

11.10.1 The Superprism Effect

Photonic crystal prisms were first studied in 1996 [126]. A couple of years later, Kosaka et al. reported anomalous refraction phenomena in 3D photonic crystals. In the experiments, a beam of light was impinged on a 3D crystal grown by self-assembly techniques. The beam angle inside the photonic crystal was observed to be highly sensitive to the wavelength and incident angle. When the wavelength changed from 1 to 0.99 μm, the beam angle swung by 50°, as sketched in Fig. 11.26 [127]. In contrast, a conventional crystal would give less than a 1° angular swing. In addition, for fixed wavelength, the refraction angle changed from −70° to 70° when the incident angle varied from –7° to 7° [125]. In contrast to conventional bulk materials, the refracted beam inside the crystal lies on the same side of the surface normal as the incident beam, a phenomenon called "negative refraction."

The direction of the refracted beam is determined by the group velocity $\mathbf{v}_g = \nabla_k \omega(\mathbf{k})$. For a given frequency ω_0, the direction of the group velocity can be obtained by plotting the equi-frequency surface $\omega(\mathbf{k}) = \omega_0$ in reciprocal space, often called the "dispersion surface" in the photonic crystal research community. At any given point $\mathbf{k}_0$, the group velocity $\mathbf{v}_g$ is parallel to the surface normal. Since the surface normal is not necessarily parallel to the $\mathbf{k}_0$, the refracted beam direction is not always parallel to the wavevector. Figure 11.27 illustrates how the superprism effect can arise. The dispersion surfaces of the incident medium (e.g., air) and the photonic crystal are plotted at two adjacent frequencies ω_1 and ω_2. In this particular case, the dispersion contour of the photonic crystal shrinks as the frequency increases while the contour of the incident medium expands. The component of the wavevector parallel to the interface (k_x) is the same on either side of the interface. For a given incident

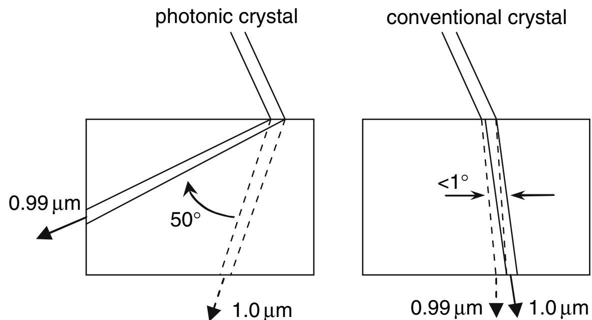


Fig. 11.26 Schematic drawing of the superprism effect observed [125,127]

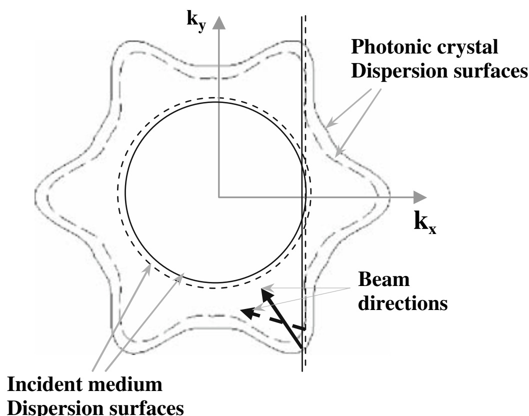


Fig. 11.27 Origin of the superprism effect. The dispersion contours at frequencies ω_1 and ω_2 ($>\omega_1$) are plotted as *lines* and *dashed lines*, respectively

angle, an increase in frequency causes the constant k_x-line to shift to the right. The slight horizontal shift of the coupling point on the two adjacent dispersion contours of the photonic crystal results in a significant change of the surface-normal direction. As illustrated in Fig. 11.27, the surface normal of the dispersion surface may turn more than 50° as the coupling point is tuned around a sharp corner, whereas the rotation angle of the wavevector could be less than 1°. This large contrast between the rotation of $\mathbf{k}$ and rotation of $\mathbf{v}_g$ is the physical origin of the superprism effect [127].

11.10.2 Transmission and Refraction of a Beam of Finite Width

In the preceding discussion, the incident wave is assumed to be a planar wave of infinite width. This, however, is never the case in reality. In this section, we discuss the refraction of a finite-width beam. We will show that in many cases, the refraction angle and transmission and reflection coefficients obtained from the planar wave case remain valid, though the physical interpretation is slightly different.

In order to quantitatively describe photonic crystal refraction phenomena, we analyze the refraction of an *ideal* rectangular beam (constant intensity across the beam width) on a photonic crystal surface [128]. Assume that the refracted beams inside the photonic crystal retain a rectangular, constant intensity envelope. For simplicity, we consider a 2D case; it is straightforward to extend the following discussion to 3D.

One can construct a contour in x–y space large enough to enclose the refraction point on the interface. The contour should cut cross the beam sufficiently far from the refraction point that all the beams separate from each other when they cross the contour. One such contour is illustrated in Fig. 11.28. For visual clarity, the reflected beams are omitted in the drawing.

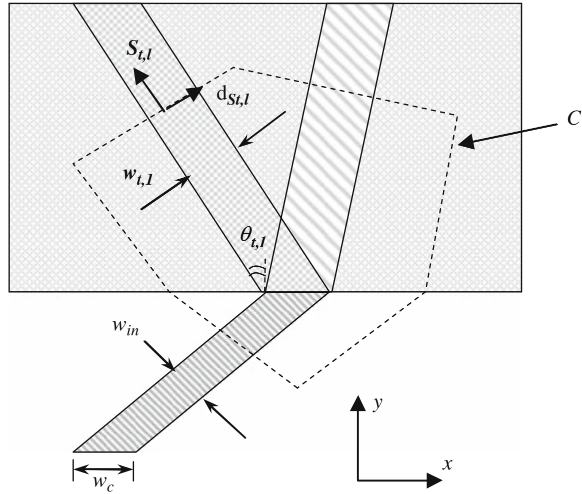


Fig. 11.28 Hypothetical perfect rectangular beam refraction

Then one can compute the following path integral along contour C:

$$0 = \int_C S\,\mathrm{d}s = -\int_{C_{\mathrm{in}}} S_{\mathrm{in}}\,\mathrm{d}s_{\mathrm{in}} + \sum_j \int_{C_{\mathrm{r},j}} S_{\mathrm{r},j}\,\mathrm{d}s_{\mathrm{r},j} + \sum_l \int_{C_{\mathrm{t},l}} S_{\mathrm{t},l}\,\mathrm{d}s_{\mathrm{t},l} \quad (11.36)$$

where S is the magnitude of the Poynting vector, ds is the line element along the contour, and the subscripts in, (r,j), and (t,l) indicate the incident beam, the jth reflected beam (not drawn), and the lth transmitted beam, respectively. The contour C is divided into segments C_{in}, $C_{\mathrm{r},j}$, and $C_{\mathrm{t},l}$ that extend across the corresponding beam width. The negative sign in front of the integral for the incident beam is attributed to the convention of using the outward surface-normal component for the surface integral; here, the incident beam has energy flowing into the contour. The integral must vanish because there is no source or absorber inside the contour.

To gain physical insight, we make the further assumption that each beam is a perfect rectangular beam. Therefore, inside each beam the Poynting vector is a constant (in the sense of cell average); outside the beams, the Poynting vector vanishes. Then Eq. (11.36) can be simplified to

$$0 = -S_{\mathrm{in}} w_{\mathrm{in}} + \sum_j S_{\mathrm{r},j} w_{\mathrm{r},j} + \sum_l S_{\mathrm{t},l} w_{\mathrm{t},l} \quad (11.37)$$

where the beam widths w_{in}, $w_{\mathrm{r},j}$, and $w_{\mathrm{t},l}$ are defined perpendicular to their respective beam propagation directions. Note the beam widths satisfy the following relations:

$$\frac{w_{\mathrm{in}}}{\cos(\theta_{\mathrm{in}})} = \frac{w_{\mathrm{t},l}}{\cos(\theta_{\mathrm{t},l})} = \frac{w_{\mathrm{r},j}}{\cos(\theta_{\mathrm{r},j})} = w_{\mathrm{c}} \quad (11.38)$$

where θ_{in}, $\theta_{\mathrm{r},j}$, and $\theta_{\mathrm{t},l}$ are the incident angle, the angle of the jth reflected beam, and that of the lth refracted beam, respectively; w_{c} is the width of any beam sectioned by the surface, as indicated in Fig. 11.28. The relation Eq. (11.38) is trivial for rectangular beams, but a rigorous proof for Gaussian beams is indeed fairly complicated. In fact, if the incident Gaussian beam width is too narrow, the refracted field in the photonic crystal may not maintain a beam form. Under such circumstances, Eq. (11.38) does not hold. In practical WDM devices, one should make every effort in design to preserve a decent beam form.

The y-component of the Poynting vector of each beam can be written as

$$(S_{\mathrm{in}})_y = S_{\mathrm{in}}\cos(\theta_{\mathrm{in}}), \qquad (S_{\mathrm{r},j})_y = -S_{\mathrm{r},j}\cos(\theta_{\mathrm{r},j}), \qquad (S_{\mathrm{t},l})_y = S_{\mathrm{t},l}\cos(\theta_{\mathrm{t},l})$$

It is straightforward to show that the conservation of energy gives rise to

$$-(S_{\mathrm{in}})_y - \sum_j (S_{\mathrm{r},j})_y + \sum_l (S_{\mathrm{t},l})_y = 0 \tag{11.39}$$

This expression can be rewritten as

$$\sum_j R_j + \sum_l T_l = \sum_j \frac{-(S_{\mathrm{r},j})_y}{(S_{\mathrm{in}})_y} + \sum_l \frac{(S_{\mathrm{t},l})_y}{(S_{\mathrm{in}})_y} = 1 \tag{11.40}$$

Here we have defined transmission and reflection coefficients R_j and T_l based on the ratios of y-components of the corresponding Poynting vectors.

To obtain Eq. (11.40), we have assumed a perfect rectangular beam. In real experiments, the laser beam is better described by a Gaussian. Consider the TM polarization of a 2D photonic crystal. Assume the incoming Gaussian beam in the homogeneous medium is given by

$$E_{\mathrm{in}}(\mathbf{x}) = \exp(\mathrm{i}\mathbf{q}_0\mathbf{x})\exp[-4x_\perp^2/w_{\mathrm{I}}^2] \tag{11.41}$$

where w_{I} is the full width of the beam at $1/e$ of the peak electric field, and $x_\perp$ is the lateral coordinate for the incident beam defined with respect to the center line. It can be proved that if w_{I} is sufficiently wide, a refracted beam in the photonic crystal retains the Gaussian-like envelope [32]:

$$E_s(\mathbf{x}) \approx t_s^{(0)} E_s^{(0)}(\mathbf{x}) \exp\left(-4x_{s\perp}^2/w_s^2\right) \tag{11.42}$$

Here the complex coupling amplitude $t_s^{(0)}$ and the mode field $E_s^{(0)}(\mathbf{x})$ for the sth mode are evaluated at $\mathbf{q}_0$ according to the theory described in Section 11.4.5.

The full width of the sth beam at $1/e$ of the peak electric field is given by w_s, and the lateral coordinate $x_{s\perp}$ is defined for the sth beam. Equations (11.38,11.39,11.40) remain valid if the Gaussian beam is sufficiently wide [129]. The above derivation assumes that the beam is sufficiently wide that the divergence of a beam is a high-order effect and can be neglected.

11.10.3 Wavelength Division Multiplexing Applications

Wavelength division multiplexing (WDM) is a bandwidth utilization technique that has had major impact on fiber-optic communications. This technique divides the transmission window of optical fibers into a large number of wavelength channels. Each channel transmits information independent of the information carried on the other channels. A key device of WDM technology is the wavelength demultiplexer illustrated in Fig. 11.29. Light of multiple wavelengths originating from an optical fiber is separated into different output waveguides through a demultiplexer. There are two main types of WDM systems: dense WDM systems having tens of channels with a wavelength spacing around 0.8 nm (100 GHz) or below and coarse WDM systems having a few wavelength channels separated by $\sim$10 nm.

It has been proposed that the photonic crystal superprism effect can be applied to wavelength demultiplexing in WDM applications [125]. The high angular dispersion of photonic crystals, of the order of 10°/nm, shows promise for separating a large number of narrowly spaced wavelengths in a small device area. Generally, highly dispersive effects are often limited to a relatively narrow bandwidth and/or accompanied by high optical loss. However, with the help of the theory presented in Section 11.4, we are able to obtain an optimized superprism demultiplexer design [32] where high dispersion of $\sim$3.5°/nm can be achieved over considerably wide bandwidth ($\sim$25 nm), sufficient for 30 wavelength channels spaced at 100 GHz. Most importantly, low optical losses ($<$3 dB) are obtained across this bandwidth, making it appealing for practical applications.

However, the beam divergence, or diffraction, in a photonic crystal is a severe issue in designing demultiplexers based on the superprism effect. A

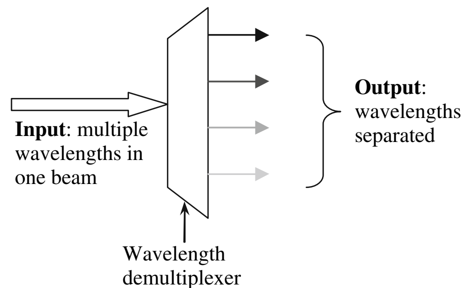


Fig. 11.29 Block diagram of a wavelength demultiplexer used in fiber-optic communications

high sensitivity to wavelength is usually accompanied with a high sensitivity to the incident angle. In a real demultiplexer device, the incident beam always has a finite width. According to the Fourier theorem, the Fourier decomposition of a beam with a finite width will always contain a continuous distribution of plane waves $e^{\mathbf{iqx}}$ whose incident wavevectors $\mathbf{q}$ are slightly different from the nominal incident wavevector $\mathbf{q}_0$. As a result, the incident angles of the incident plane-wave components will have a distribution as well. In principle, we can increase the beam width to make Δq sufficiently narrow. On the other hand, in designing a photonic crystal wavelength demultiplexer, it is always desirable to make the device smaller. A wide beam demands a wide input surface of the photonic crystal. In addition, the spacing of the adjacent output waveguides must be larger than the beam width. For a large number of output waveguides, the device size will increase significantly.

The relationship between wavelength resolution limit and device size has been discussed by Baba et al. [130]. Consider an incident beam of width w_0, the incident angles of decomposed plane waves are distributed over an angular width of $\Delta\alpha = 2\lambda/(n\pi w_0)$. The angular distribution of refracted wavevectors in the photonic crystal has a width $\Delta\theta = (\partial\theta/\partial\alpha)_\omega \Delta\alpha$, where the partial derivative is taken at constant frequency. The wavelength resolution is given by

$$\begin{aligned}\frac{\Delta\lambda}{\lambda} &= \frac{1}{\lambda}\left(\frac{\partial\lambda}{\partial\theta}\right)_\alpha \Delta\theta = \frac{1}{\lambda}\left(\frac{\partial\lambda}{\partial\theta}\right)_\alpha \left(\frac{\partial\theta}{\partial\alpha}\right)_\omega \Delta\alpha \\ &= \frac{2}{n\pi w_0}\left(\frac{\partial\lambda}{\partial\theta}\right)_\alpha \left(\frac{\partial\theta}{\partial\alpha}\right)_\omega\end{aligned}$$

Introducing the normalized frequency a/λ, the wavelength resolution may also be expressed as

$$\frac{\Delta\lambda}{\lambda} = \frac{2\lambda^2}{n\pi w_0 a}\left(\frac{\partial(a/\lambda)}{\partial\theta}\right)_\alpha \left(\frac{\partial\theta}{\partial\alpha}\right)_\omega = \frac{2\lambda^2}{n\pi w_0 a}\cdot\frac{p}{q} \tag{11.43}$$

where $p = (\partial\theta/\partial\alpha)_\omega$ and $q = [\partial\theta/\partial(a/\lambda)]_\alpha$. The wavelength resolutions for a hexagonal photonic crystal lattice were numerically calculated through Eq. (11.43). It was found that $q/p > 75$ is possible for such a lattice. For an incident beam having $w_0 = 115\,\mu\text{m}$, a superprism demultiplexer of size $(6.5\,\text{cm})^2$ is capable of separating 56 wavelengths spaced 0.4 nm apart.

Another interesting aspect of the wavelength resolution is discussed in the literature [130]. In an ordinary medium, the far-field Gaussian beam divergence relation $w = L\cdot\Delta\theta$ is valid only if $L > \pi n w_0^2/\lambda$. If we assume that this equation changes to $L > (\pi n w_0^2/\lambda)p$ for photonic crystal refraction, then there could be a limitation on the length of superprism demultiplexer. However, further analysis is needed to clarify the details of this limit.

One major difficulty of designing a superprism-based wavelength demultiplexer arises from the low crosstalk requirement for WDM applications. The propagation length L required to limit the crosstalk to a value X is given by [131],

$$L(X) = z_0 K(X)/[\eta - H(X)] \quad (11.44)$$

where $z_0 = 4\lambda/(\pi n_e \Delta\theta^2)$ is the Rayleigh range for the beam inside a photonic crystal (n_e is an effective index), η is the ratio of the angular separation of adjacent channels to the individual beam divergence angle $\Delta\theta$, $K(X)$ and $H(X)$ are two parameters depending on the crosstalk X only. From this relation, the design of superprism-based wavelength demultiplexers was systematically examined to show the size advantage of the photonic crystal approach [131]. Two figures of merit that represent the device size and wavelength resolution were introduced. Particularly, in the case of equal frequency separation, the device area was found to grow with the number of wavelength channels, N, as N^4. It was estimated that a 16-channel superprism demultiplexer with about 5 nm wavelength resolution would occupy an area of 0.22 mm^2.

A number of experiments have been conducted to study the wavelength separation capability of the superprism effect. One interesting design employed a semi-circular photonic crystal slab, an input waveguide pointing toward the center of the semi-circle, and output waveguides extending radially outward from the circumference [132]. The superprism effect was also demonstrated in low-index contrast 3D polymeric photonic crystals [56]. Wavelength sensitivities were experimentally measured in 2D photonic crystals [133,134]. These experiments corroborated that the beam direction in a photonic crystal does change sensitively with wavelength. Moreover, it is possible to design a photonic crystal such that the negative refraction phenomenon and the superprism effect occur for the same wavelength range and coupling conditions. Thus, negative refraction was also exploited to compensate the beam divergence in the photonic crystal, resulting in the demonstration of a 4-channel wavelength demultiplexer with a channel spacing of 8 nm and a crosstalk level of –6.5 dB or better [135].

11.10.4 Application in Electro-optics, Nonlinear Optics, and Sensing

Electro-optic (EO) control of the superprism effect has been discussed in [136]. In general, the refractive index of a material changes with an applied field as

$$\Delta n_{ij} = -(1/2) n_{ij}^3 (r_{ijk} E_k + s_{ijkl} E_k E_l)$$

where r_{ijk} and s_{ijkl} are the linear and quadratic EO coefficients. The shifted dielectric constants are given by

$$\varepsilon_{ij}^{*} = \varepsilon_0 (n_{ij} + \Delta n_{ij})^2 \tag{11.45}$$

The ferroelectric material lead lanthanum zirconate titanate (PLZT) has $s_{3333}{\sim}4 \times 10^{-16}\ \mathrm{m}^2/\mathrm{V}^2$. An applied field of 6 V/μm gives a change in the dielectric constant of around 0.12. It is estimated that a field strength of 6 V/μm would be sufficient to deflect the beam by about 49° in a 2D PLZT photonic crystal.

The superprism effect has also been investigated in nonlinear photonic crystals [137]. A pump beam propagating in a 2D photonic crystal can be used to change the refractive index, shifting the direction of a refracted signal beam. Because the pump beam and the signal beam have two distinct wavelengths, it is possible to design their dispersion characteristics separately. The change in the refractive index due to the pump beam is given by

$$\delta n(\mathbf{x}) = \varepsilon_0 c n n_2 |\mathbf{E}(\mathbf{x})|^2/2 \tag{11.46}$$

Using this relation, a self-consistent calculation gives the photonic band structure of the pumped photonic crystal. In addition, the dispersion surface at a given pump power level can be computed. Note that the pump power (or Poynting vector) is proportional to the group velocity:

$$P \sim < \varepsilon(\mathbf{x})|\mathbf{E}(\mathbf{x})|^2 > v_g \tag{11.47}$$

where the brackets denote the spatial average over a unit cell. For a given pump power, a lower group velocity results in a large field $|\mathbf{E}(\mathbf{x})|^2$, and therefore a larger δn according to Eq. (11.46). By choosing the pump wavelength close to the band edge where the group velocity is small, it is possible to significantly enhance the nonlinear effect.

Nonlinear photonic crystals also exhibit self-induced superprism effects [137]. In this case, no pump beam is needed; the modification of the refractive index is owing to the power of the signal beam itself. For a superprism made of GaAs, whose Kerr coefficient is assumed to be $n_2 = 3 \times 10^{16}\ \mathrm{m}^2/\mathrm{W}$, it is estimated that about a few GW/cm^2 are needed to observe tens of degrees of beam deflection. A further example takes into account the beam width and finite thickness of a 2D photonic crystal slab. For an optical beam 10 μm wide, a deflection of 10° can be achieved at 1.55 μm by varying the optical power from 1.3 to 3.63 W in a 0.25 μm thick GaAs 2D photonic crystal slab. It is predicted that InSb, ZnSe, or polymeric materials whose Kerr coefficients are more than two orders of magnitude larger may give even smaller switching power.

The sensitivity of the beam direction to the refractive index change could also be used in sensing applications. Macroporous photonic crystals formed from a colloidal crystal template were theoretically investigated [138]. Such a porous structure is particularly conducive for sensing applications because the analyte can be adsorbed onto the large surface areas inside the photonic crystal.

Adsorption of an analyte changes the polymer refractive index, shifting the beam direction. It is predicted that a 70° beam angle change can be obtained for 0.63% change in the polymer refractive index.

11.10.5 Phase Prism

In the preceding discussion, the high sensitivity of beam direction refers to the beam inside the photonic crystal. However, for a number of applications, such as agile steering of laser beams, it is desired that the direction of an output beam exiting a photonic crystal have a high sensitivity to the wavelength, incident angle, or refractive index. If the photonic crystal is a slab whose input and output surfaces are parallel to each other as shown in Fig. 11.30(a), the output beam direction is given by Snell's law,

$$n_3 \sin\phi = n_1 \sin\alpha \tag{11.48}$$

owing to the conservation of the surface tangential wavevector component. Therefore, no matter how sensitive the beam direction is inside the photonic crystal, the output beam is fixed at an angle independent of the wavelength. Here we have assumed that there is only one output beam, which is equivalent to the zeroth order diffraction of the output surface. This assumption is consistent with most cases encountered in simulation and experiments.

However, if the output surface is not parallel to the input surface, the output angle is no longer governed by Eq. (11.48). Consider a simple case where the output surface is perpendicular to the input surface [139] as sketched in Fig. 11.30(a). Here the photonic crystal has a square lattice and the principal axes of the crystal are rotated 45° with respect to the input surface. The dispersion surface of the photonic crystal is sketched in Fig. 11.30(b). The circle represents the dispersion surface of the output medium (identical to the input medium in this case). If the incident angle increases slightly, as indicated by the thicker black arrow in Fig. 11.30(a), the constant k_x-lines for the input coupling shift slightly to the right, as shown in Fig. 11.30(b). The coupling point on the dispersion surface moves from a to b. The output coupling is governed by the constant k_y-lines, which shift significantly because $|k_{y,a}-k_{y,b}| >> |k_{x,a}-k_{x,b}|$. Correspondingly, the

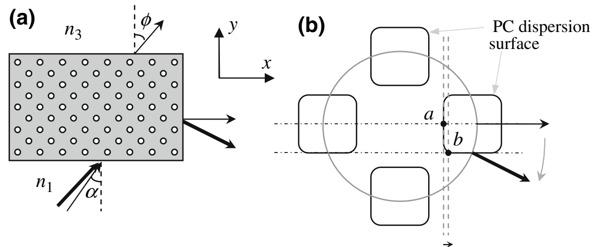


Fig. 11.30 A phase prism. (**a**) Device schematic; (**b**) dispersion surface

output beam direction changes from horizontal to substantially downward. The output surface does not necessarily need to be perpendicular to the input surface to achieve a sensitive output angle. A 45° output surface has also been investigated [140]. Because the output beam direction in the homogeneous medium is determined by the phase velocity or wavevector k, rather than the group velocity, this effect is sometimes called a phase prism or k-prism. Note that although gratings may achieve similar sensitivity effects for their output angles, the sensitivity is much smaller at angles where the optical loss is acceptable.

11.11 Negative Refraction

Light striking an interface between air and a material with constant refractive index n is refracted at an angle determined by Snell's law. Naturally occurring, homogeneous materials generally have a positive refractive index. As a result, the incident and refracted beams lie on opposite sides of the surface normal. However, for a material with $n < 0$, called a "negative index material," the direction of refraction is reversed, leading to novel optical behavior [141]. For example, negative index materials may allow "perfect" lenses, with image sharpness below the Rayleigh resolution limit [142]. Such "superlenses" could have profound impact on resolution-limited technologies such as photolithography. One approach to creating artificial negative index materials is to use subwavelength, metallic elements with a resonant response to light. These structures, known as "metamaterials," exhibit negative refraction at microwave frequencies [143]. A thin slab of silver approximates some of the properties of a negative index material [142,144], but with appreciable optical loss at optical frequencies. It turns out that photonic crystals can exhibit negative refraction in the optical frequency range [145,146,147] with relatively low loss. We have already seen one example of negative refraction above, in Fig. 11.26.

A photonic crystal lens employing negative refraction is shown schematically in Fig. 11.31(a). A source on one side of the lens produces an image on the far side. An optical beam impinging on the surface of a photonic crystal may split into a number of beams owing to the presence of different branches of the dispersion surface at a given frequency. In addition, a beam exiting a photonic crystal surface may split into a number of beams owing to the diffraction effect. Image formation requires that these issues be solved. A set of conditions sufficient to guarantee single-beam all-angle negative refraction was given [146] as follows: (1) The constant-frequency contour of the photonic crystal is all convex with a negative photonic effective mass; (2) all incoming wavevectors at such a frequency are included within the constant-frequency contour of the photonic crystal; (3) the frequency is below $\pi c/a_s$, where a_s is the surface-parallel periodicity of the photonic crystal. It was shown that such a set of conditions could be satisfied by a 2D square lattice photonic crystal at a certain frequency in the first photonic band for the TE polarization. In Fig. 11.31(b), we illustrate

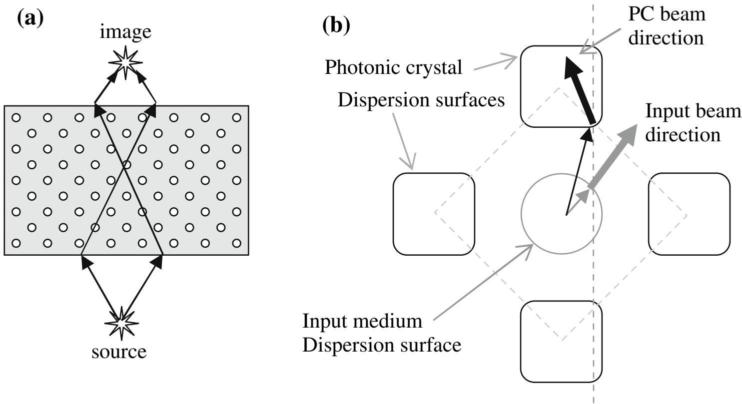


Fig. 11.31 Focus through a photonic crystal flat-lens. (**a**) Ray analysis showing a ray-crossing inside the photonic crystal lens; (**b**) dispersion surface analysis in reciprocal space. The *gray dashed square* delineates the Brillouin zone, which is rotated 45° because the input surface has Miller indices (11.11). The *vertical dashed line*, to which the ends of the wavevectors in the input medium and photonic crystal align, indicates the conservation of the surface tangential wavevector component

a situation where all three criteria are met. Here the input medium could, for example, be air. Generally, the first condition can be met if

$$\omega - \omega_0 = (k_x - k_{x0})^2/m^* + (k_y - k_{y0})^2/m^* \tag{11.49}$$

where the "effective mass" $m^* < 0$ (note that m^* does not actually have dimensions of mass; the terminology is adopted from solid-state physics). A negative effective mass means that the contour shrinks as the frequency ω increases. Therefore, the group velocity points inward for each contour near $\mathbf{k}_0$ as illustrated in Fig. 11.31(b). The second condition is satisfied if the equi-frequency contour of the incoming medium has a width (in the direction parallel to the surface) narrower than the dispersion surface contour of the photonic crystal. The contours drawn in Fig. 11.31(b) clearly satisfy this condition. As the width of the input medium dispersion surface becomes narrower, a beam incident at any angle will have its surface tangential wavevector component contained in the width of the top photonic crystal dispersion contour. Thus all-angle negative refraction is achieved. The last condition can usually be satisfied by operating in the lowest frequency band of a photonic crystal.

In addition, surface termination and thickness must be optimized to reduce internal reflection. The formation of an image with resolution at or below the wavelength is demonstrated numerically in Refs. [146,148]. Note that for the photonic crystal superlens to form a real image on the other side of the slab as shown in Fig. 11.31(a), the slab must be thick enough for the ray-crossing to be located inside the photonic crystal. For this reason, a thick slab may be required

to focus a distant object. Issues such as aberration and relative phase between the point-object and image are discussed in Ref. [146].

Microwave experiments were conducted to verify negative refraction in a millimeter-scale square lattice formed of alumina rods [149]. An incident Gaussian beam shifts to the "wrong" direction after passing the photonic crystal slab, giving an effective negative index of refraction -1.94. In contrast, a slab made of polystyrene pellets shifts the output beam as normal, giving a positive index of 1.46. Both numbers are in good agreement with simulation. Subsequent experiments demonstrated all-angle negative refraction at optical telecommunications wavelengths [150].

Some interesting applications of all-angle negative refraction can be found in structureless confinement of light in optical devices, as demonstrated experimentally in Ref. [151]. Negative refraction may further lead to flexible and efficient waveguiding and optical routing.

For several years after Pendry proposed the "perfect lens," [142] it was suggested that negative refraction might violate causality. However direct electromagnetic simulations have shown that this is not the case, at least in a photonic crystal [147]. The system studied was a 2D photonic crystal that consists of a hexagonal lattice of dielectric rods with $\varepsilon = 12.96$ and $r = 0.35a$. A Gaussian beam in vacuum is incident upon the photonic crystal surface at 30°, as shown in Fig. 11.32(a). The incident beam has a normalized frequency $\omega a/2\pi c = 0.58$, at which the effective index of refraction of the photonic crystal is $n = -0.7$. One key discovery of this transient study is that the Gaussian wave, once entering the photonic crystal, is trapped near the surface for a long time. According to Fig. 11.32(b)–(d), it takes approximately $45T$ ($T = 2\pi/\omega$) for the refracted wave to reorganize itself and eventually propagate along the negative direction. This time is much longer than the time difference, $3T$, for the outer ray to catch up to the inner ray of the Gaussian beam upon arrival at the surface. Therefore, intuitively, the photonic crystal waits long enough to adjust to the impact of the slightly later arrival of the outer ray, such that the final field structure near the surface remains a causal effect of both inner and outer rays. Neither causality nor the speed of light is violated according to this simulation.

We note that for an inhomogeneous/anisotropic system like a photonic crystal, a distinction can be made between left-handed materials and negative index materials [146,147]. A photonic crystal exhibits left-handed (LH) behavior if a refracted mode satisfies $\langle \mathbf{S} \rangle \cdot \mathbf{k} < 0$, where $\langle \mathbf{S} \rangle$ is the time-averaged Poynting vector and $\mathbf{k}$ is the reduced wavevector of this mode [147]. This condition is a direct generalization of the LH condition, $(\mathbf{E} \times \mathbf{H}) \cdot \mathbf{k} < 0$, for a homogeneous, isotropic medium. It implies that the angle between $\langle \mathbf{S} \rangle$ and $\mathbf{k}$ must be greater than 90°. On the other hand, the condition for negative refraction requires only that the $\langle \mathbf{S} \rangle$ is on the other side of the surface normal, or $\langle S_x \rangle \cdot k_x < 0$ if the surface lies along the x-axis. There exist refracted modes that exhibit negative refraction while being right-handed. In all preceding analysis, we have assumed the wavevector $\mathbf{k}$ is in the first Brillouin zone. However, we should be careful in applying the above criteria to modes in higher photonic

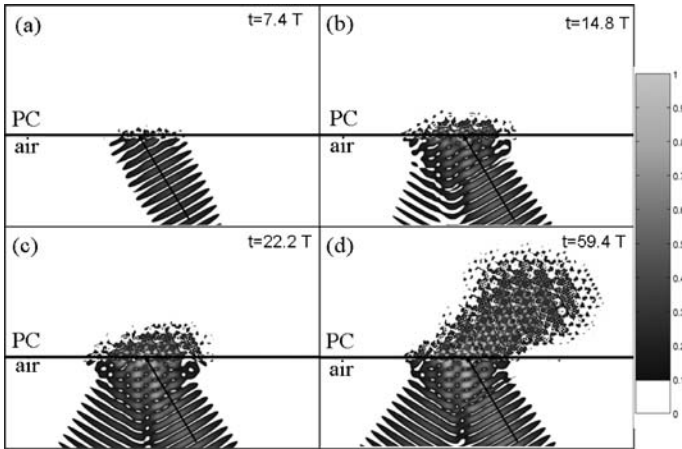


Fig. 11.32 Time domain field evolution for negative refraction on a photonic crystal surface. (**a**) An incident wave at 30° with the surface normal enters the photonic crystal. (**b**) The wave does not presume a certain direction and seems "undecisive" where to go. (**c**) The wave re-arranges itself inside the photonic crystal. (**d**) It finally propagates in the negative – in respect to the surface normal – direction. (Illustration courtesy of Costas Soukolis' group)

bands where the Fourier components in the first Brillouin zone are relatively weak compared to the components in other Brillouin zones.

Although negative refraction has aroused widespread interest, it remains an open question whether it can overcome the aperture limitations imposed on all imaging systems. Any practical photonic crystal superlens has a finite lateral extent, i.e., a finite aperture. When such a finite aperture is present, according to the principles of Fourier optics, certain information whose spatial frequency is much higher than $2\pi/d$ (d is the aperture size) is lost behind the aperture. The lost information cannot be regenerated by using any medium, negative or positive.

11.12 Concluding Remarks

Photonic crystal structures have been introduced in a wide range of optoelectronic devices, as reviewed in this chapter. In some of these devices, such as lasers and modulators, obvious advantages such as small size and lower power consumption of photonic crystal-based devices have been demonstrated. Some remaining key challenges of individual devices have been mentioned in respective sections. Device research will continue to pose scientific questions and help further our understanding of photonics and solid-state physics (e.g., an ideal crystal surface coupling theory for both Miller-indexed surfaces and quasi-periodic surfaces presented in Section 11.4, also see a review [152]). At the same time, we may see the results of laboratory device research being adopted into real-world applications in the near future. In the last decade, optical loss (both propagating loss and coupling loss) of photonic crystal waveguides has been

significantly reduced to a practical level. Furthermore, most photonic crystal device structures made using electron-beam lithography are also amenable to deep ultraviolet (DUV) lithography. This makes photonic crystals amenable to cost-effective VLSI fabrication technology, through which photonic crystal devices may be integrated with optical and electronic systems. For example, in optical communications, photonic crystals may provide a compact structural platform for important devices in fiber-to-the-home (FTTH) systems. Some of the devices used in FTTH systems, such as lasers, modulators, wavelength multiplexers, and optical filters, have been discussed in detail in this chapter.

In addition, photonic crystals may provide alternatives for solving nano- and giga-challenges in electronics. As transistors on computer chips shrink toward a nanometer in dimension and run faster than a gigahertz in speed, technological challenges are emerging to thwart the continual improvement in microprocessors that computer users have enjoyed over the last few decades. While many novel nanoelectronic approaches aim to overcome the challenges of dimensional shrinkage, the speed of computer chips is no longer accelerating as fast as before. Among the issues are the time delay on metal wires and other interconnection bottlenecks on computer chips. On-chip data transmission at 40 GHz or above will be difficult to accommodate within the current electrical transmission architecture. Recent Pentium chips already have 50% of total power dissipated in interconnection rather than transistor switching, exacerbating the overheating of chips. Optical interconnects that can transmit data through modulated laser signals in optical waveguides may solve these problems, consuming less power and providing higher data transmission speed. Photonic crystal lasers and modulators, discussed in this chapter, may be key devices for on-chip data transmission. The promise of building low-power-threshold lasers and the power advantage of silicon-based photonic crystal modulators [153] continue to drive photonic crystal research toward addressing giga-challenges. Photonic crystals, as an optical structure platform, also help reduce the dimensions of optical interconnect components, saving the precious estate on a silicon chip.

Last but certainly not least, photonic crystal negative index materials may lead to alternative lithographic techniques for nanoscale patterning of next-generation computer chips.

Acknowledgment W. Jiang thanks the Air Force Office of Scientific Research (Dr. Gernot Pomrenke), Air Force Research Laboratory (Dr. Robert L. Nelson), and NASA for support during the period of writing.

References

1. E. Yablonovitch, Inhibited spontaneous emission in solid-state physics and electronics, *Phys. Rev. Lett.* **58**, 2059–2062 (1987).
2. S. John, Strong localization of photons in certain disordered dielectric superlattices, *Phys. Rev. Lett.* **58**, 2486–2489 (1987).

3. W. M. Robertson, G. Arjavalingam, R. D. Meade, K. D. Brommer, A. M. Rappe and J. D. Joannopoulos, Measurement of photonic band-structure in a 2-dimensional periodic dielectric array, *Phys. Rev. Lett.* **68**, 2023–2026 (1992).
4. D. F. Sievenpiper, M. E. Sickmiller and E. Yablonovitch, 3D wire mesh photonic crystals, *Phys. Rev. Lett.* **76**, 2480–2483 (1996).
5. J. D. Joannopoulos, R. D. Meade and J. N. Winn, *Photonic Crystals: Molding the Flow of Light* (Princeton University Press, Princeton, 1995).
6. S. G. Johnson and J. D. Joannopoulos, Block-iterative frequency-domain methods for Maxwell's equations in a planewave basis, *Opt. Express* **8**, 173–190 (2001).
7. K. M. Ho, C. T. Chan, C. M. Soukoulis, R. Biswas and M. Sigalas, Photonic band-gaps in 3-dimensions - new layer-by-layer periodic structures, *Solid State Commun.* **89**, 413–416 (1994).
8. S. G. Johnson and J. D. Joannopoulos, Three-dimensionally periodic dielectric layered structure with omnidirectional photonic band gap, *Appl. Phys. Lett.* **77**, 3490–3492 (2000).
9. M. L. Povinelli, S. G. Johnson, S. H. Fan and J. D. Joannopoulos, Emulation of two-dimensional photonic crystal defect modes in a photonic crystal with a three-dimensional photonic band gap, *Phys. Rev. B* **64**, 075313 (2001).
10. S. G. Johnson, S. H. Fan, P. R. Villeneuve, J. D. Joannopoulos and L. A. Kolodziejski, Guided modes in photonic crystal slabs, *Phys. Rev. B* **60**, 5751–5758 (1999).
11. S. G. Johnson, P. R. Villeneuve, S. H. Fan and J. D. Joannopoulos, Linear waveguides in photonic-crystal slabs, *Phys. Rev. B* **62**, 8212–8222 (2000).
12. W. T. Lau and S. H. Fan, Creating large bandwidth line defects by embedding dielectric waveguides into photonic crystal slabs, *Appl. Phys. Lett.* **81**, 3915–3917 (2002).
13. M. L. Povinelli, S. G. Johnson, E. Lidorikis, J. D. Joannopoulos and M. Soljacic, Effect of a photonic band gap on scattering from waveguide disorder, *Appl. Phys. Lett.* **84**, 3639–3641 (2004).
14. S. G. Johnson, M. L. Povinelli, M. Soljacic, A. Karalis, S. Jacobs and J. D. Joannopoulos, Roughness losses and volume-current methods in photonic-crystal waveguides, *Appl. Phys. B-Lasers Opt.* **81**, 283–293 (2005).
15. N. Stefanou and A. Modinos, Impurity bands in photonic insulators, *Phys. Rev. B* **57**, 12127–12133 (1998).
16. A. Yariv, Y. Xu, R. K. Lee and A. Scherer, Coupled-resonator optical waveguide: a proposal and analysis, *Opt. Lett.* **24**, 711–713 (1999).
17. T. Baba, N. Fukaya and A. Motegi, Clear correspondence between theoretical and experimental light propagation characteristics in photonic crystal waveguides, *Electron. Lett.* **37**, 761–762 (2001).
18. M. Notomi, A. Shinya, S. Mitsugi, E. Kuramochi and H. Y. Ryu, Waveguides, resonators and their coupled elements in photonic crystal slabs, *Opt. Express* **12**, 1551–1561 (2004).
19. K. S. Kunz and R. J. Luebbers, *The Finite-Difference Time-Domain Method for Electromagnetics* (CRC Press, Boca Raton, 1993).
20. A. Taflove and S. C. Hagness, *Computational Electrodynamics: The Finite-Difference Time-Domain Method* (Artech House, Norwood, 2005).
21. S. Y. Shi, C. H. Chen and D. W. Prather, Revised plane wave method for dispersive material and its application to band structure calculations of photonic crystal slabs, *Appl. Phys. Lett.* **86**, 043104 (2005).
22. M. Qiu, Effective index method for heterostructure-slab-waveguide-based two-dimensional photonic crystals, *Appl. Phys. Lett.* **81**, 1163–1165 (2002).
23. P. R. Villeneuve, S. Fan, S. G. Johnson and J. D. Joannopoulos, Three-dimensional photon confinement in photonic crystals of low-dimensional periodicity, *IEE Proceedings–Optoelectronics* **145**, 384–390 (1998).
24. S. G. Johnson, S. Fan, A. Mekis and J. D. Joannopoulos, Multipole-cancellation mechanism for high-Q cavities in the absence of a complete photonic band gap, *Appl. Phys. Lett.* **78**, 3388–3390 (2001).

25. J. Vuckovic, M. Loncar, H. Mabuchi and A. Scherer, Optimization of the Q factor in photonic crystal microcavities, *IEEE J. Quantum Electron.* **38**, 850–856 (2002).
26. K. Srinivasan and O. Painter, Momentum space design of high-Q photonic crystal optical cavities, *Opt. Express* **10**, 670–684 (2002).
27. H. Y. Ryu, M. Notomi and Y. H. Lee, High-quality-factor and small-mode-volume hexapole modes in photonic-crystal-slab nanocavities, *Appl. Phys. Lett.* **83**, 4294–4296 (2003).
28. B. S. Song, S. Noda, T. Asano and Y. Akahane, Ultra-high-Q photonic double-heterostructure nanocavity, *Nat. Mater.* **4**, 207–210 (2005).
29. E. Kuramochi, M. Notomi, S. Mitsugi, A. Shinya, T. Tanabe and T. Watanabe, Ultrahigh-Q photonic crystal nanocavities realized by the local width modulation of a line defect, *Appl. Phys. Lett.* **88**, 041112 (2006).
30. T. Asano, B. S. Song, Y. Akahane and S. Noda, Ultrahigh-Q nanocavities in two-dimensional photonic crystal slabs, *IEEE J. Sel. Top. Quantum Electron.* **12**, 1123–1134 (2006).
31. T. Tanabe, M. Notomi, E. Kuramochi, A. Shinya and H. Taniyama, Trapping and delaying photons for one nanosecond in an ultrasmall high-Q photonic-crystal nanocavity, *Nat. Photonics* **1**, 49–52 (2007).
32. W. Jiang, R. T. Chen and X. J. Lu, Theory of light refraction at the surface of a photonic crystal, *Phys. Rev. B* **71**, 245115 (2005).
33. R. D. Meade, K. D. Brommer, A. M. Rappe and J. D. Joannopoulos, Electromagnetic bloch waves at the surface of a photonic crystal, *Phys. Rev. B* **44**, 10961–10964 (1991).7
34. Z. Y. Li and K. M. Ho, Light propagation in semi-infinite photonic crystals and related waveguide structures, *Phys. Rev. B* **68**, 155101 (2003).
35. T. Ochiai and J. Sanchez-Dehesa, Superprism effect in opal-based photonic crystals, *Phys. Rev. B* **64**, 245113 (2001).
36. X. N. Chen, W. Jiang, J. Q. Chen and R. T. Chen, Theoretical study of light refraction in three -dimensional photonic crystals, *J. Lightwave Technol.* **25**, 2469–2474 (2007).
37. J. B. Pendry and A. Mackinnon, Calculation of photon dispersion-relations, *Phys. Rev. Lett.* **69**, 2772–2775 (1992).
38. K. Ohtaka, T. Ueta and K. Amemiya, Calculation of photonic bands using vector cylindrical waves and reflectivity of light for an array of dielectric rods, *Phys. Rev. B* **57**, 2550–2568 (1998).
39. N. Stefanou, V. Karathanos and A. Modinos, Scattering of electromagnetic-waves by periodic structures, *J. Phys.-Condes. Matter* **4**, 7389–7400 (1992).
40. J. Bravo-Abad, T. Ochiai and J. Sanchez-Dehesa, Anomalous refractive properties of a two-dimensional photonic band-gap prism, *Phys. Rev. B* **67**, 115116 (2003).
41. K. Sakoda, Symmetry, degeneracy, and uncoupled modes in 2-dimensional photonic lattices, *Phys. Rev. B* **52**, 7982–7986 (1995).
42. K. Sakoda, Transmittance and bragg reflectivity of 2-dimensional photonic lattices, *Phys. Rev. B* **52**, 8992–9002 (1995).
43. P. Bienstman and R. Baets, Optical modelling of photonic crystals and VCSELs using eigenmode expansion and perfectly matched layers, *Opt. Quantum Electron.* **33**, 327–341 (2001).
44. W. Jiang and R. T. Chen, Rigorous analysis of diffraction gratings of arbitrary profiles using virtual photonic crystals, *J. Opt. Soc. Am. A-Opt. Image Sci. Vis.* **23**, 2192–2197 (2006).
45. M. G. Moharam and T. K. Gaylord, Diffraction analysis of dielectric surface-relief gratings, *J. Opt. Soc. Am.* **72**, 1385–1392 (1982).
46. T. Baba and M. Nakamura, Photonic crystal light deflection devices using the superprism effect, *IEEE J. Quantum Electron.* **38**, 909–914 (2002).
47. E. Yablonovitch, T. J. Gmitter and K. M. Leung, Photonic band-structure - the face-centered-cubic case employing nonspherical atoms, *Phys. Rev. Lett.* **67**, 2295–2298 (1991).

48. T. F. Krauss, R. M. Delarue and S. Brand, Two-dimensional photonic-bandgap structures operating at near infrared wavelengths, *Nature* **383**, 699–702 (1996).
49. D. Labilloy, H. Benisty, C. Weisbuch, T. F. Krauss, R. M. Delarue, V. Bardinal, R. Houdre, U. Oesterle, D. Cassagne and C. Jouanin, Quantitative measurement of transmission, reflection, and diffraction of two-dimensional photonic band gap structures at near-infrared wavelengths, *Phys. Rev. Lett.* **79**, 4147–4150 (1997).
50. S. Y. Lin, J. G. Fleming, D. L. Hetherington, B. K. Smith, R. Biswas, K. M. Ho, M. M. Sigalas, W. Zubrzycki, S. R. Kurtz and J. Bur, A three-dimensional photonic crystal operating at infrared wavelengths, *Nature* **394**, 251–253 (1998).
51. M. J. Escuti and G. P. Crawford, Holographic photonic crystals, *Opt. Eng.* **43**, 1973–1987 (2004).
52. L. J. Wu, Y. C. Zhong, C. T. Chan, K. S. Wong and G. P. Wang, Fabrication of large area two- and three-dimensional polymer photonic crystals using single refracting prism holographic lithography, *Appl. Phys. Lett.* **86**, 241102 (2005).
53. J. Q. Chen, W. Jiang, X. N. Chen, L. Wang, S. S. Zhang and R. T. Chen, Holographic three-dimensional polymeric photonic crystals operating in the 1550 nm window, *Appl. Phys. Lett.* **90**, 93102 (2007).
54. J. Koch, F. Korte, C. Fallnich, A. Ostendorf and B. N. Chichkov, Direct-write subwavelength structuring with femtosecond laser pulses, *Opt. Eng.* **44**, 051103-5 (2005).
55. J. Serbin and M. Gu, Experimental evidence for superprism effects in three-dimensional polymer photonic crystals, *Adv. Mater.* **18**, 221–224 (2006).
56. J. Serbin and M. Gu, Superprism phenomena in waveguide-coupled woodpile structures fabricated by two-photon polymerization, *Opt. Express* **14**, 3563–3568 (2006).
57. A. Imhof and D. J. Pine, Ordered macroporous materials by emulsion templating, *Nature* **389**, 948–951 (1997).
58. O. D. Velev, P. M. Tessier, A. M. Lenhoff and E. W. Kaler, Materials – A class of porous metallic nanostructures, *Nature* **401**, 548–548 (1999).
59. A. Imhof, W. L. Vos, R. Sprik and A. Lagendijk, Large dispersive effects near the band edges of photonic crystals, *Phys. Rev. Lett.* **83**, 2942–2945 (1999).
60. Y. A. Vlasov, X. Z. Bo, J. C. Sturm and D. J. Norris, On-chip natural assembly of silicon photonic bandgap crystals, *Nature* **414**, 289–293 (2001).
61. Z. L. Wang, C. T. Chan, W. Y. Zhang, N. B. Ming and P. Sheng, Three-dimensional self-assembly of metal nanoparticles: Possible photonic crystal with a complete gap below the plasma frequency, *Phys. Rev. B* **64**, 113108 (2001).
62. O. D. Velev and E. W. Kaler, Structured porous materials via colloidal crystal templating: From inorganic oxides to metals, *Adv. Mater.* **12**, 531–534 (2000).
63. M. Diop and R. A. Lessard, Fabrication techniques of high quality photonic crystals, *Optical Interconnects and VLSI Photonics, 2004 Digest of the LEOS Summer Topical Meetings*, pp. 79–80 (2004).
64. J. F. Bertone, P. Jiang, K. S. Hwang, D. M. Mittleman and V. L. Colvin, Thickness dependence of the optical properties of ordered silica-air and air-polymer photonic crystals, *Phys. Rev. Lett.* **83**, 300–303 (1999).
65. L. Wang, W. Jiang, X. Chen, L. Gu, J. Chen and R. T. Chen, Fabrication of polymer photonic crystal superprism structures using polydimethylsiloxane soft molds, *J. Appl. Phys.* **101**, 114316-6 (2007).
66. M. H. Qi, E. Lidorikis, P. T. Rakich, S. G. Johnson, J. D. Joannopoulos, E. P. Ippen and H. I. Smith, A three-dimensional optical photonic crystal with designed point defects, *Nature* **429**, 538–542 (2004).
67. F. Garcia-Santamaria, M. J. Xu, V. Lousse, S. H. Fan, P. V. Braun and J. A. Lewis, A germanium inverse woodpile structure with a large photonic band gap, *Adv. Mater.* **19**, 1567–1570 (2007).
68. S. Kawakami, Fabrication of submicrometre 3D periodic structures composed of Si/Si Osub 2., *Electron. Lett.* **33**, 1260–1261 (1997).

69. S. Noda, K. Tomoda, N. Yamamoto and A. Chutinan, Full three-dimensional photonic bandgap crystals at near-infrared wavelengths, *Science* **289**, 604–606 (2000).
70. M. Settle, M. Salib, A. Michaeli and T. F. Krauss, Low loss silicon on insulator photonic crystal waveguides made by 193 nm optical lithography, *Opt. Express* **14**, 2440–2445 (2006).
71. M. H. Shih, W. J. Kim, W. Kuang, J. R. Cao, H. Yukawa, S. J. Choi, J. D. O'brien, P. D. Dapkus and W. K. Marshall, Two-dimensional photonic crystal Mach–Zehnder interferometers, *Appl. Phys. Lett.* **84**, 460–462 (2004).
72. H. G. Park, S. H. Kim, S. H. Kwon, Y. G. Ju, J. K. Yang, J. H. Baek, S. B. Kim and Y. H. Lee, Electrically driven single-cell photonic crystal laser, *Science* **305**, 1444–1447 (2004).
73. E. Purcell, Spontaneous emission probabilities at radio frequencies, *Phys. Rev.* **69**, 681 (1946).
74. J. M. Gerard and B. Gayral, Strong Purcell effect for InAs quantum boxes in three-dimensional solid-state microcavities, *J. Lightwave Technol.* **17**, 2089–2095 (1999).
75. O. Painter, R. K. Lee, A. Scherer, A. Yariv, J. D. O'Brien, P. D. Dapkus and I. Kim, Two-dimensional photonic band-gap defect mode laser, *Science* **284**, 1819–1821 (1999).
76. H. Altug, D. Englund and J. Vuckovic, Ultrafast photonic crystal nanocavity laser, *Nat. Phys.* **2**, 484–488 (2006).
77. S. P. Ogawa, M. Imada, S. Yoshimoto, M. Okano and S. Noda, Control of light emission by 3D photonic crystals, *Science* **305**, 227–229 (2004).
78. W. D. Zhou, J. Sabarinathan, B. Kochman, E. Berg, O. Qasaimeh, S. Pang and P. Bhattacharya, Electrically injected single-defect photon bandgap surface-emitting laser at room temperature, *Electron. Lett.* **36**, 1541–1542 (2000).
79. W. D. Zhou, J. Sabarinathan, P. Bhattacharya, B. Kochman, E. W. Berg, P. C. Yu and S. W. Pang, Characteristics of a photonic bandgap single defect microcavity electroluminescent device, *IEEE J. Quantum Electron.* **37**, 1153–1160 (2001).
80. H. G. Park, J. K. Hwang, J. Huh, H. Y. Ryu, Y. H. Lee and J. S. Kim, Nondegenerate monopole-mode two-dimensional photonic band gap laser, *Appl. Phys. Lett.* **79**, 3032–3034 (2001).
81. K. Sakoda, *Optical Properties of Photonic Crystals* (Springer, Berlin, 2001).
82. A. Mekis, M. Meier, A. Dodabalapur, R. E. Slusher and J. D. Joannopoulos, Lasing mechanism in two-dimensional photonic crystal lasers, *Appl. Phys. A* **69**, 111–114 (1999).
83. J. P. Dowling, M. Scalora, M. J. Bloemer and C. M. Bowden, The Photonic band-edge laser – a new approach to gain enhancement, *J. Appl. Phys.* **75**, 1896–1899 (1994).
84. M. Boroditsky, T. F. Krauss, R. Coccioli, R. Vrijen, R. Bhat and E. Yablonovitch, Light extraction from optically pumped light-emitting diode by thin-slab photonic crystals, *Appl. Phys. Lett.* **75**, 1036–1038 (1999).
85. S. H. Fan, P. R. Villeneuve, J. D. Joannopoulos and E. F. Schubert, High extraction efficiency of spontaneous emission from slabs of photonic crystals, *Phys. Rev. Lett.* **78**, 3294–3297 (1997).
86. A. A. Erchak, D. J. Ripin, S. Fan, P. Rakich, J. D. Joannopoulos, E. P. Ippen, G. S. Petrich and L. A. Kolodziejski, Enhanced coupling to vertical radiation using a two-dimensional photonic crystal in a semiconductor light-emitting diode, *Appl. Phys. Lett.* **78**, 563–565 (2001).
87. H. Ichikawa and T. Baba, Efficiency enhancement in a light-emitting diode with a two-dimensional surface grating photonic crystal, *Appl. Phys. Lett.* **84**, 457–459 (2004).
88. A. J. Danner, J. J. Raftery, P. O. Leisher and K. D. Choquette, Single mode photonic crystal vertical cavity lasers, *Appl. Phys. Lett.* **88**, 091114 (2006).
89. S. H. Fan, P. R. Villeneuve and J. D. Joannopoulos, Channel drop tunneling through localized states, *Phys. Rev. Lett.* **80**, 960–963 (1998).
90. A. Sharkawy, S. Y. Shi and D. W. Prather, Multichannel wavelength division multiplexing with photonic crystals, *Appl. Optics* **40**, 2247–2252 (2001).

91. H. A. Haus, *Waves and Fields in Optoelectronics* (Prentice-Hall, Englewood Cliffs, 1984).
92. Y. Akahane, T. Asano, B. S. Song and S. Noda, Fine-tuned high-Q photonic-crystal nanocavity, *Opt. Express* **13**, 1202–1214 (2005).
93. S. Noda, A. Chutinan and M. Imada, Trapping and emission of photons by a single defect in a photonic bandgap structure, *Nature* **407**, 608–610 (2000).
94. A. Chutinan, M. Mochizuki, M. Imada and S. Noda, Surface-emitting channel drop filters using single defects in two-dimensional photonic crystal slabs, *Appl. Phys. Lett.* **79**, 2690–2692 (2001).
95. Y. Akahane, T. Asano, B. S. Song and S. Noda, Investigation of high-Q channel drop filters using donor-type defects in two-dimensional photonic crystal slabs, *Appl. Phys. Lett.* **83**, 1512–1514 (2003).
96. B. S. Song, S. Noda and T. Asano, Photonic devices based on in-plane hetero photonic crystals, *Science* **300**, 1537–1537 (2003).
97. T. Asano, W. Kunishi, M. Nakamura, B. S. Song and S. Noda, Dynamic wavelength tuning of channel-drop device in two-dimensional photonic crystal slab, *Electron. Lett.* **41**, 37–38 (2005).
98. C. Manolatou, M. J. Khan, S. H. Fan, P. R. Villeneuve, H. A. Haus and J. D. Joannopoulos, Coupling of modes analysis of resonant channel add-drop filters, *IEEE J. Quantum Electron.* **35**, 1322–1331 (1999).
99. K. H. Hwang and G. H. Song, Design of a high-Q channel add-drop multiplexer based on the two-dimensional photonic-crystal membrane structure, *Opt. Express* **13**, 1948–1957 (2005).
100. Z. Zhang and M. Qiu, Compact in-plane channel drop filter design using a single cavity with two degenerate modes in 2D photonic crystal slabs, *Opt. Express* **13**, 2596–2604 (2005).
101. H. Takano, B. S. Song, T. Asano and S. Noda, Highly efficient in-plane channel drop filter in a two-dimensional heterophotonic crystal, *Appl. Phys. Lett.* **86**, 241101 (2005).
102. Z. Y. Zhang and M. Qiu, Coupled-mode analysis of a resonant channel drop filter using waveguides with mirror boundaries, *J. Opt. Soc. Am. B-Opt. Phys.* **23**, 104–113 (2006).
103. H. Takano, B. S. Song, T. Asano and S. Noda, Highly efficient multi-channel drop filter in a two-dimensional hetero photonic crystal, *Opt. Express* **14**, 3491–3496 (2006).
104. A. Shinya, S. Mitsugi, E. Kuramochi and M. Notomi, Ultrasmall multi-channel resonant-tunneling filter using mode gap of width-tuned photonic-crystal waveguide, *Opt. Express* **13**, 4202–4209 (2005).
105. W. Jiang and R. T. Chen, Multichannel optical add-drop processes in symmetrical waveguide-resonator systems, *Phys. Rev. Lett.* **91**, 213901 (2003).
106. M. Notomi, K. Yamada, A. Shinya, J. Takahashi, C. Takahashi and I. Yokohama, Extremely large group-velocity dispersion of line-defect waveguides in photonic crystal slabs, *Phys. Rev. Lett.* **87**, 253902 (2001).
107. H. Gersen, T. J. Karle, R. J. P. Engelen, W. Bogaerts, J. P. Korterik, N. F. Van Hulst, T. F. Krauss and L. Kuipers, Real-space observation of ultraslow light in photonic crystal waveguides, *Phys. Rev. Lett.* **94**, 073903 (2005).
108. M. Soljacic, S. G. Johnson, S. H. Fan, M. Ibanescu, E. Ippen and J. D. Joannopoulos, Photonic-crystal slow-light enhancement of nonlinear phase sensitivity, *J. Opt. Soc. Am. B* **19**, 2052–2059 (2002).
109. Y. A. Vlasov, M. O'Boyle, H. F. Hamann and S. J. Mcnab, Active control of slow light on a chip with photonic crystal waveguides, *Nature* **438**, 65–69 (2005).
110. Y. Q. Jiang, W. Jiang, L. L. Gu, X. N. Chen and R. T. Chen, 80-micron interaction length silicon photonic crystal waveguide modulator, *Appl. Phys. Lett.* **87**, 221105 (2005).
111. L. L. Gu, W. Jiang, X. N. Chen, L. Wang and R. T. Chen, High speed silicon photonic crystal waveguide modulator for low voltage operation, *Appl. Phys. Lett.* **90**, 071105 (2007).
112. A. S. Liu, R. Jones, L. Liao, D. Samara-Rubio, D. Rubin, O. Cohen, R. Nicolaescu and M. Paniccia, A high-speed silicon optical modulator based on a metal-oxide-semiconductor capacitor, *Nature* **427**, 615–618 (2004).

113. T. Chu, H. Yamada, S. Ishida and Y. Arakawa, Thermooptic switch based on photonic-crystal line-defect waveguides, *IEEE Photonics Technol. Lett*. **17**, 2083–2085 (2005).
114. C. A. Barrios, V. R. Almeida, R. Panepucci and M. Lipson, Electrooptic modulation of silicon-on-insulator submicrometer-size waveguide devices, *J. Lightwave Technol*. **21**, 2332–2339 (2003).
115. C. A. Barrios, V. R. De Almeida and M. Lipson, Low-power-consumption short-length and high-modulation-depth silicon electrooptic modulator, *J. Lightwave Technol*. **21**, 1089–1098 (2003).
116. Q. F. Xu, B. Schmidt, S. Pradhan and M. Lipson, Micrometre-scale silicon electro-optic modulator, *Nature* **435**, 325–327 (2005).
117. R. A. Soref and B. R. Bennett, Electrooptical effects in silicon, *IEEE J. Quantum Electron*. **23**, 123–129 (1987).
118. G. V. Treyz, P. G. May and J. M. Halbout, Silicon Mach–Zehnder wave-guide interferometers based on the plasma dispersion effect, *Appl. Phys. Lett*. **59**, 771–773 (1991).
119. C. Z. Zhao, G. Z. Li, E. K. Liu, Y. Gao and X. D. Liu, Silicon-on-insulator Mach–Zehnder wave-guide interferometers operating at 1.3 Mu-M, *Appl. Phys. Lett*. **67**, 2448–2449 (1995).
120. R. C. Alferness, Waveguide electrooptic modulators, *IEEE Trans. Microw. Theory Tech*. **30**, 1121–1137 (1982).
121. D. A. B. Miller, Rationale and challenges for optical interconnects to electrical chips, *Proc. IEEE* **88**, 728–749 (2000).
122. R. T. Chen, L. Lin, C. Choi, Y. J. Liu, B. Bihari, L. Wu, S. Tang, R. Wickman, B. Picor, M. K. Hibbs-Brenner, J. Bristow and Y. S. Liu, Fully embedded board-level guided-wave optoelectronic interconnects, *Proc. IEEE* **88**, 780–793 (2000).
123. E. A. Camargo, H. M. H. Chong and R. M. De La Rue, 2D Photonic crystal thermo-optic switch based on AlGaAs/GaAs epitaxial structure, *Opt. Express* **12**, 588–592 (2004).
124. L. L. Gu, W. Jiang, X. N. Chen and R. T. Chen, Thermooptically tuned photonic crystal waveguide silicon-on-insulator Mach–Zehnder interferometers, *IEEE Photonics Technol. Lett*. **19**, 342–344 (2007).
125. H. Kosaka, T. Kawashima, A. Tomita, M. Notomi, T. Tamamura, T. Sato and S. Kawakami, Superprism phenomena in photonic crystals, *Phys. Rev. B* **58**, 10096–10099 (1998).
126. S. Y. Lin, V. M. Hietala, L. Wang and E. D. Jones, Highly dispersive photonic band-gap prism, *Opt. Lett*. **21**, 1771–1773 (1996).
127. H. Kosaka, T. Kawashima, A. Tomita, M. Notomi, T. Tamamura, T. Sato and S. Kawakami, Superprism phenomena in photonic crystals: Toward microscale lightwave circuits, *J. Lightwave Technol*. **17**, 2032–2038 (1999).
128. W. Jiang, C. Tian, Y. Jiang, Y. Chen, X. Lu and R. T. Chen, Superprism effect and light refraction and propagation in photonic crystals, *Proc. SPIE*, **5733**, 50–57 (2005).
129. W. Jiang, unpublished (2004).
130. T. Baba and T. Matsumoto, Resolution of photonic crystal superprism, *Appl. Phys. Lett*. **81**, 2325–2327 (2002).
131. B. Momeni and A. Adibi, Systematic design of superprism-based photonic crystal demultiplexers, *IEEE J. Sel. Areas Commun*. **23**, 1355–1364 (2005).
132. L. J. Wu, M. Mazilu, T. Karle and T. F. Krauss, Superprism phenomena in planar photonic crystals, *IEEE J. Quantum Electron*. **38**, 915–918 (2002).
133. J. J. Baumberg, N. M. B. Perney, M. C. Netti, M. D. C. Charlton, M. Zoorob and G. J. Parker, Visible-wavelength super-refraction in photonic crystal superprisms, *Appl. Phys. Lett*. **85**, 354–356 (2004).
134. A. Lupu, E. Cassan, S. Laval, L. El Melhaoui, P. Lyan and J. M. Fedeli, Experimental evidence for superprism phenomena in SOI photonic crystals, *Opt. Express* **12**, 5690–5696 (2004).

135. B. Momeni, J. D. Huang, M. Soltani, M. Askari, S. Mohammadi, M. Rakhshandehroo and A. Adibi, Compact wavelength demultiplexing using focusing negative index photonic crystal superprisms, *Opt. Express* **14**, 2413–2422 (2006).
136. D. Scrymgeour, N. Malkova, S. Kim and V. Gopalan, Electro-optic control of the superprism effect in photonic crystals, *Appl. Phys. Lett.* **82**, 3176–3178 (2003).
137. N. C. Panoiu, M. Bahl and R. M. Osgood, Optically tunable superprism effect in nonlinear photonic crystals, *Opt. Lett.* **28**, 2503–2505 (2003).
138. T. Prasad, V. Colvin and D. Mittleman, Superprism phenomenon in three-dimensional macroporous polymer photonic crystals, *Phys. Rev. B* **67**, 165103 (2003).
139. C. Y. Luo, M. Soljacic and J. D. Joannopoulos, Superprism effect based on phase velocities, *Opt. Lett.* **29**, 745–747 (2004).
140. T. Baba, T. Matsumoto and M. Echizen, Finite difference time domain study of high efficiency photonic crystal superprisms, *Opt. Express* **12**, 4608–4613 (2004).
141. V. G. Veselago, The Electrodynamics of substances with simultaneously negative values of ε and μ, *Sov. Phys. Usp.* **10**, 509–514 (1968).
142. J. B. Pendry, Negative refraction makes a perfect lens, *Phys. Rev. Lett.* **85**, 3966–3969 (2000).
143. R. A. Shelby, D. R. Smith and S. Schultz, Experimental verification of a negative index of refraction, *Science* **292**, 77–79 (2001).
144. N. Fang, H. Lee, C. Sun and X. Zhang, Sub-diffraction-limited optical imaging with a silver superlens, *Science* **308**, 534–537 (2005).
145. M. Notomi, Theory of light propagation in strongly modulated photonic crystals: Refractionlike behavior in the vicinity of the photonic band gap, *Phys. Rev. B* **62**, 10696–10705 (2000).
146. C. Luo, S. G. Johnson, J. D. Joannopoulos and J. B. Pendry, All-angle negative refraction without negative effective index, *Phys. Rev. B* **65**, 201104 (2002).
147. S. Foteinopoulou, E. N. Economou and C. M. Soukoulis, Refraction in media with a negative refractive index, *Phys. Rev. Lett.* **90**, 107402 (2003).
148. C. Y. Luo, S. G. Johnson, J. D. Joannopoulos and J. B. Pendry, Subwavelength imaging in photonic crystals, *Phys. Rev. B* **68**, 045115 (2003).
149. E. Cubukcu, K. Aydin, E. Ozbay, S. Foteinopoulou and C. M. Soukoulis, Negative refraction by photonic crystals, *Nature* **423**, 604–605 (2003).
150. A. Berrier, M. Mulot, M. Swillo, M. Qiu, L. Thylen, A. Talneau and S. Anand, Negative refraction at infrared wavelengths in a two-dimensional photonic crystal, *Phys. Rev. Lett.* **93**, 073902 (2004).
151. D. W. Prather, S. Y. Shi, D. M. Pustai, C. H. Chen, S. Venkataraman, A. Sharkawy, G. J. Schneider and J. Murakowski, Dispersion-based optical routing in photonic crystals, *Opt. Lett.* **29**, 50–52 (2004).
152. E. Istrate and E. H. Sargent, Photonic crystal heterostructures and interfaces, *Rev. Mod. Phys.* **78**, 455 (2006).
153. W. Jiang, L. Gu, X. Chen, R. T. Chen, Photonic crystal waveguide modulators for silicon photonics: Device physics and some recent progress, *Solid State Electronics*, **51**, 1278 (2007).

Chapter 12
Two-Photon Polymerization – High Resolution 3D Laser Technology and Its Applications

Aleksandr Ovsianikov and Boris N. Chichkov

Abstract The development of high-precision fabrication techniques is an essential factor and a driving power for the increasing progress in the field of nanotechnology. The femtosecond (10^{-15} s) laser technology opens a broad range of opportunities for cost-efficient manufacturing with high resolution and unprecedented flexibility. In this chapter we discuss principles and advances in two-photon activated laser processing.

12.1 Introduction

The history of the development of the femtosecond laser systems is an intense and exciting one. The associated scientific findings have produced a leap in the understanding of many physical and chemical processes. Due to the extremely short duration of the pulse, the peak power of modern commercially available tabletop laser can exceed 200 GW [1]. In comparison, the total power of power stations of Germany, including solar power plants and wind power stations, are expected to approach 120 GW by the year 2010 [2]. Application of femtosecond pulses allows the introduction of laser energy with unprecedented precision in space and time. First short pulses, demonstrated in the mid-1970 s, have been produced by the dye lasers. Despite many disadvantages of the dye lasers, they have been used to demonstrate a number of exciting results in the variety of scientific research areas, by making it possible to study various processes at a femtosecond time scale. By applying the light source that has the same time scale as molecular vibrations, one could observe and even control the outcome of chemical and biological reactions in real time. The application of the femtosecond lasers to probe molecular dynamics has been explored for more than two decades now and it was recognized by the award of the Nobel Prize in Chemistry in 1999 [3]. The high peak powers also permit efficient wavelength conversion using nonlinear crystals, thus broadening the areas of investigation of

B.N. Chichkov
Nanotechnology Department, Laser Zentrum Hannover e.V., Hannover, Germany
e-mail: b.chichkov@lzh.de

A. Korkin, F. Rosei (eds.), *Nanoelectronics and Photonics*,
DOI: 10.1007/978-0-387-76499-3_12, © Springer Science+Business Media, LLC 2008

highly nonlinear processes in atomic, molecular, plasma, and solid-state physics, and to access previously unexplored states of matter. In addition, the ultrashort x-ray pulses, generated from the plasma produced by femtosecond laser, have been applied to probe short- and long-range atomic dynamics. Nowadays, femtosecond lasers are almost routinely used for high-resolution imaging, such as multi-photon microscopy based on a highly localized laser radiation interaction with the sample.

Theoretical model for the multi-photon absorption (MPA) was developed in 1931 by Maria Göppert-Meyer [4], three decades prior to its experimental observation [5]. The probability of n-photon absorption is proportional to the nth power of the photon flux density; consequently high photon flux densities are required in order to observe this phenomenon. In fact, MPA was one of the first effects demonstrated with the help of lasers, since intensities much higher than provided by other light sources could be achieved. It was demonstrated that an atom can absorb two or more photons simultaneously, thus allowing electron transition to the states that cannot be reached with a single-photon absorption. Atom excitation with both single-photon absorption and two-photon absorption (TPA) are compared schematically in Fig. 12.1a,b. TPA is mediated by a virtual state (dashed line in Fig. 12.1b), which has an extremely short lifetime (several femtoseconds). Thus, TPA is only possible if a second photon is absorbed before the decay of this virtual state. Note that excited energy levels S_1 and S_2, shown in Fig. 12.1a,b, are not exactly the same, since the selection rules for single-photon and two-photon absorption are different [6]. This fact also implies that MPA can reveal information about transitions not accessible by one-photon processes. Since the probability of the TPA is proportional to the square of the intensity of the laser radiation, favourable conditions for TPA in the first place are created in the focus of the laser beam (see Fig. 12.1c). Thus, the interaction region is strongly confined. The main

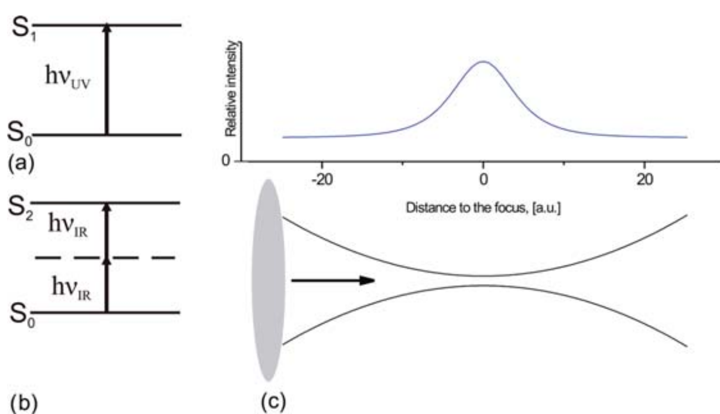


Fig. 12.1 Excitation through a (**a**) single-photon and (**b**) two-photon absorption; (**c**) intensity distribution along the propagation direction of the focused laser beam

advantages of multi-photon microscopy are high spatial resolution and the ability to selectively excite specific molecules. Using multi-photon microscopy one can, for example, observe the spatial distribution of one specific molecule inside a living cell and create 3D images with submicrometre resolution.

Short pulses with modest energy can provide huge peak powers. This makes femtosecond pulses very suitable for laser ablation of materials; the resulting cuts are much cleaner because the laser pulses turn the material into plasma rather than melting it. Since the interaction is very short, even such easily destroyable substances as living or biological materials can be ablated without changing the properties of the remaining material. MPA also allows to pattern surfaces or to induce changes within the materials transparent at the wavelength of the applied laser radiation. This method is used to write 3D waveguides and microfluidic channels inside various types of glasses. Currently, the structural size of such directly written patterns is of the order of few micrometres.

The main focus of this tutorial chapter is the two-photon polymerization (2 PP) technique and its applications. This microstructuring technique is based on the interaction of femtosecond laser radiation with a photosensitive material through MPA, which induces a highly localized chemical reaction leading to polymerization of the photosensitive material. Current capability of the 2 PP technique allows to create arbitrary 3D structures with resolution down to 100 nm. Great flexibility provided by this technique and a vast variety of processable materials make it useful for a large number of applications. 2 PP is still a very young technology and has a great potential for further improvements.

The development of femtosecond lasers progressed considerably in the last three decades. In the early 1990 s commercially available solid-state femtosecond lasers based on Kerr lens mode-locking were introduced, making high powers and short pulses available from systems fitting on a standard optical table laboratory setup. The central emission wavelength of a Ti:sapphire-based system can be adjusted in the range of 700–1000 nm. Other groups, working on the Ti:sapphire-based systems, developed alternative concepts employing saturable absorption effects in semiconductors instead of the Kerr lens effect. The self-starting and more reliable mode-locked operation are listed among the comparative advantages of these laser concepts. Among others, continuous wave or pulsed laser systems based on active solid-state media such as Nd:glass, Nd:YVO4, and Yb:YVO4, historically used for industrial applications, have been utilized for ultrashort pulse generation. These lasers emit at wavelengths above 1000 nm and are pumped by comparably cheap laser diodes. There has also been considerable development in the femtosecond fibre-laser systems. In this case fibres doped with active elements play the role of a lasing media and a resonator simultaneously. These systems are extremely compact and robust. In addition, these fibre-lasers can be pumped by conventional laser diodes used in telecommunication, resulting in high reliability and low costs.

Modern femtosecond laser is a computer-controlled turnkey system. Cost reduction and ease of operation are important steps towards the

industrialization of the femtosecond technology. The resolution of microstructuring techniques utilizing femtosecond laser radiation continues to improve rapidly, demonstrating a great potential for the applications in the field of nanotechnology.

This chapter describes the basic principles, main advantages, and some applications of the 2 PP technique. Section 12.2 provides some insights into the chemical and physical processes undermining the photoinduced polymerization process. A few exciting examples of the structures fabricated by the 2 PP technique are demonstrated. Section 12.3 opens with the basic introduction into the femtosecond pulse generation; it is devoted to the detailed description of the experimental setup and main materials currently applied for 2 PP microstructuring. Section 12.4 describes different applications of 2 PP technique explored by our group.

12.2 Two-Photon Polymerization

Two-photon polymerization (2 PP) is a direct laser writing technique, which allows the fabrication of 3D structures with a resolution (structure size) down to 100 nm [7, 8]. While material photosensitive in the UV range (350–400 nm) is usually transparent to the applied near-infrared (780–800 nm Ti:sapphire) laser radiation, only two-photon absorption in the small focus area can initiate polymerization process. This technique, which will be described below in detail, allows the fabrication of computer-generated 3D structures by direct laser "recording" into the volume of a photosensitive material. Due to the threshold behaviour and nonlinear nature of the 2 PP process, resolution beyond the diffraction limit can be realized by controlling the laser pulse energy and the number of applied pulses. Figure 12.2 shows three scanning electron microscope images of 3D microstructures fabricated by the 2 PP technique. One can see the strength of this technology and envision many potential applications.

Stereolithography, which is a rapid 3D prototyping process, and the 2 PP technology are based on a similar mechanism – light triggers a chemical reaction, leading to polymerization of a photosensitive material. Polymerization is a process in which monomers or weakly cross-linked polymers (liquid or solid)

Fig. 12.2 Scanning electron microscope images of 3D structures fabricated by 2 PP technique

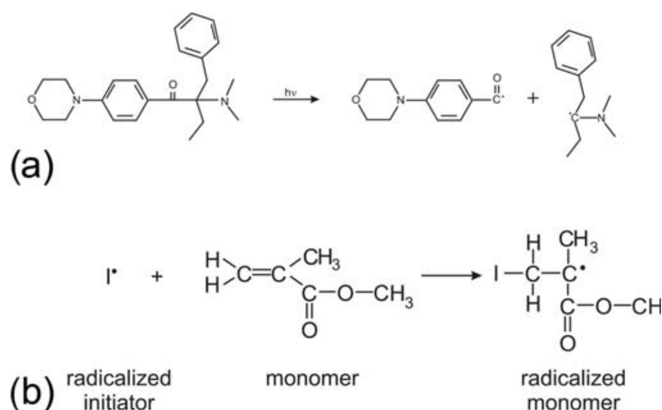


Fig. 12.3 (**a**) Cleavage of a photoinitiator, 2-benzyl-2-dimethylamino-4′-morpholinobutyrophenone, results from UV photon absorption; (**b**) radical polymerization of methyl methacrylate (MMA): I′ is an initial radical or an intermediate in the reaction chain

interconnect and form 3D network of highly cross-linked polymer (solid). Photoinitiators – molecules which have low photodissociation energy – are often added in order to increase the material photosensitivity. Absorption of a photon leading to a bond cleavage (photodissociation) and formation of highly reactive radicals is illustrated in Figure 12.3a on the example of 2-benzyl-2-dimethylamino-4′-morpholinobutyrophenone (Irgacure369, Ciba SC, Switzerland). Absorption of a UV photon breaks C – C bond and results in the formation of two radicals, which react with the monomer, e.g. methyl methacrylate, and initiate radical polymerization (Fig. 12.3b). The reaction is terminated when the two radicals react with each other.

In stereolithography, a UV laser, applied to scan the surface of the photosensitive material, produces 2D patterns of polymerized material (Fig. 12.4a). The UV laser radiation induces photopolymerization through single-photon absorption at the surface of the material. Therefore, with stereolithography it is only possible to fabricate 3D structures using a layer-by-layer approach. Since photosensitive materials are usually transparent in the infrared and highly absorptive in the UV range, one can initiate two-photon polymerization with IR laser pulses within the small volume of the material by precisely focused near-infrared femtosecond laser pulses. Figure 12.4 provides a simplified illustration of the difference between single-photon and two-photon activated processing. A material is polymerized along the trace of the moving laser focus, thus enabling fabrication of any desired polymeric 3D pattern by direct "recording" into the volume of photosensitive material. In a subsequent processing step the material which was not exposed to the laser radiation, and therefore stayed unpolymerized, is removed and the fabricated structure is revealed. The material sensitive in the UV range (λ_{UV}) can be polymerized by irradiation with the infrared light of approximately double wavelength

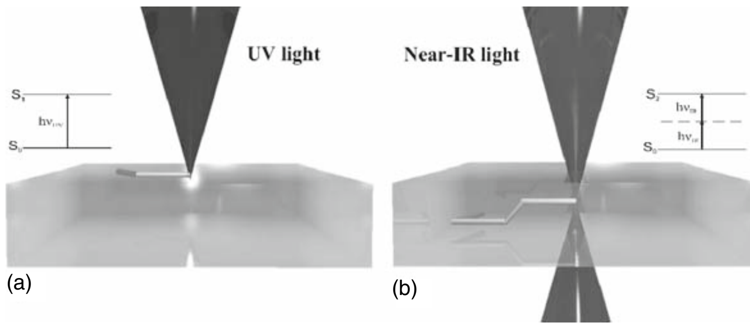


Fig. 12.4 Photosensitive material processing by (**a**) a single-photon absorption with UV light. Light is absorbed at the surface of the photosensitive material. Two-dimensional patterns can be produced by photopolymerization; (**b**) two-photon absorption with near-infrared light. TPA and following chemical reactions are confined in the focal volume, and the rest of the laser radiation passes through the material without interaction. Respective *insets* in the figures illustrate (**a**) single-photon absorption; (**b**) two-photon absorption processes

($\lambda_{IR} = 2\lambda_{UV}$), under the condition that the intensity of the radiation is high enough to initiate TPA.

Since femtosecond lasers provide very high peak intensities at the moderate average laser power, they present a very suitable light source creating favourable conditions for TPA and are commonly used for 2 PP technique.

The resolution of stereolithography depends on the size of the focal spot and is limited by diffraction, thus the minimum feature size cannot be smaller than half of the applied laser wavelength. In reality, due to technical reasons inherent to this technology, the lateral resolution of stereolithography is in the range of a few micrometres [9]. Since TPA is nonlinear and displays threshold behaviour, structural resolution beyond the diffraction limit can be realized. Structures with feature size down to 100 nm (and even better) have been demonstrated by several groups, which is almost an order of magnitude smaller than the laser wavelength (800 nm)! A voxel (volume pixel), which has the shape of an ellipsoid, can be seen as a basic unit structure (building block) polymerized in the focal volume by irradiation of the photosensitive material with laser. Intensity threshold for polymerization is defined as the minimum laser radiation intensity required for the initiation of the polymerization process leading to an irreversible change in the material. 2 PP is an accumulative process, which might result from the absorption of a number of pulses, and not necessarily of a single laser pulse. Finally, the size of the single voxel depends on the irradiation dose. Figure 12.5 illustrates the structural resolution as a function of the light intensity. Since only the area of the laser focus volume, where the intensity exceeds the polymerization threshold, will contribute to the 2 PP process, the resolution can be tuned by adjusting the pulse energy and the number of applied pulses. Theoretically, infinitely small structures can be produced by 2 PP technique. In reality, the main limiting factors are fluctuations in the laser pulse energies,

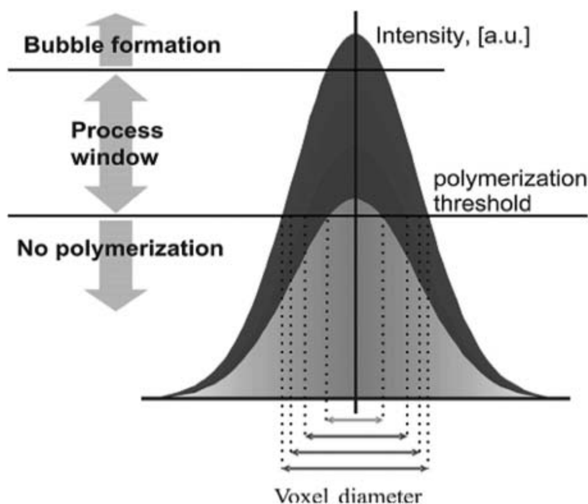


Fig. 12.5 Intensity distribution of the laser radiation at the cross-section of the focal spot. Only the part of the focal volume, where the intensity exceeds the threshold intensity for polymerization, contributes to the 2 PP process, and thus defines the resolution

limited pointing stability of the laser, and the positioning system performance. As the 2 PP technology is getting more advanced, the size of the material building blocks will become an important factor.

12.3 Materials and Methods

12.3.1 Generation of Ultrashort Laser Pulses

As was discussed in the introduction, due to the short pulse duration and the potential for strong focusing, optical pulses can be used to generate extremely high optical intensities even at moderate average laser powers. A conventional femtosecond oscillator output provides a train of pulses that appear at high repetition rates of the order of 100 MHz (or less), therefore the output is said to be "quasi-continuous". The main distinction is that in a continuous wave (CW) laser operation, all of the longitudinal modes are oscillating with random phases, i.e. they are out of phase at any time and the output signal in the time domain is comparable to that of a light bulb (see Fig. 12.6a). Fixing the phase relation between the single longitudinal modes will result in periodically appearing moments when all of the modes are oscillating in phase (Fig. 12.6b), therefore it is said that the modes are locked. The mode-locking modifies the output signal of the laser to a train of pulses, this way very short pulses can be generated.

Conventional Ti:sapphire-based femtosecond laser oscillator can support both CW and mode-lock regimes, the trick is to make the "short-pulse" regime more advantageous. Mode-locking is achieved by the modulation of the losses in the resonator. For this purpose a saturable absorber, an optical component with a certain optical loss, which is reduced for high optical intensities, is

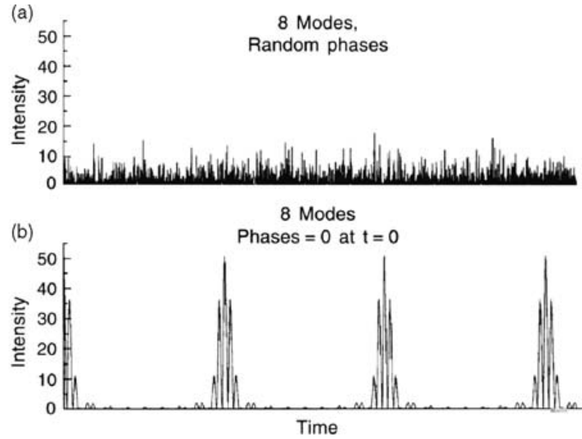


Fig. 12.6 Laser output simulation for eight modes: **(a)** random phase–continuous wave (CW) operation; **(b)** mode-locked operation

introduced. Every time a pulse hits the saturable absorber, it will saturate the absorption and temporarily reduce the losses. Therefore, the saturable absorber suppresses weaker pulses as well as any continuous background light and as a consequence the pulsed regime is made more advantageous. In addition, the saturable absorber attenuates the leading wing of the circulating pulse and tends to decrease the pulse duration. Alternatively, the Kerr lens effect in the Ti:sapphire crystal, which also introduces intensity-dependent effects, can be used. A Kerr medium exhibits intensity-dependent refractive index, thus every time the pulse passes through such medium it will be focused. By decreasing the width of the slit, placed at the output of the Kerr medium, one creates conditions where less-focused lower intensity pulses will experience higher losses (Fig. 12.7). In general, Ti:sapphire lasers operate in the CW mode at the start, but exhibit significant fluctuations of the laser power. In each resonator round trip, the saturable absorber creates favourable conditions for the light which has somewhat higher intensities, since this light can saturate the absorption slightly more than light with lower intensities. After many round trips, a single pulse will remain. However, such self-starting is not always achieved; often the initial intensity fluctuation is achieved artificially by jerking one of the optical elements (e.g. prism).

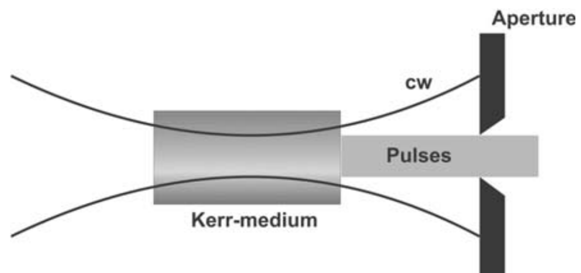


Fig. 12.7 The principle of Kerr lens mode-locking

12.3.2 Experimental Setup for 2PP

The main factors determining the performance of the 2 PP system are the sample positioning precision, the laser system stability, and the flexibility of the scanning algorithm. The schematic representation of the experimental setup is shown in Fig. 12.8. The femtosecond solid-state Ti:sapphire laser generates pulses with duration of 120 fs and a repetition rate of 94 MHz. The central emission wavelength of such a laser can be tuned between 700 and 1000 nm. Unless otherwise noted, for all the experiments described in this tutorial, laser radiation at the central emission wavelength of 780 nm was applied. A small portion of the light exiting the laser is guided into the spectrum analyser for continuous monitoring of the laser emission spectrum. The λ/2-plate mounted on the computer-controlled rotational stage is used to rotate the polarization of the laser beam. In combination with the polarization-sensitive beam-splitter, it enables continuous adjustment of the average power of the beam entering the AOM (acousto-optic modulator). The AOM is adjusted such that the first diffraction order of the beam can pass the diaphragm aperture, while the zero order is blocked. By controlling the AOM on/off state with computer-generated TTL signal, it is used as a laser shutter. In order to completely fill the aperture of a focusing optic and to achieve optimal focusing conditions, the beam is expanded to a diameter of about 10 mm by a telescope. A highly sensitive CCD camera is mounted behind the last dichroic mirror to provide online process observation. The refractive index of the polymer is slightly changed by a 2 PP process, and the polymerized patterns become visible immediately. The relative position of the laser focus within the sample is controlled by two galvo-scanner mirrors (angular range ±12.5°, resolution 6.7 μrad) and three linear translational stages (*xyz*, resolution 10 nm, maximal travel distance 2.5 cm). For the fabrication of structures presented in Fig. 12.2 and in the latter sections, unless otherwise noted, a 100× microscope objective lens (Zeiss, Plan Apochromat, NA = 1.4) was used to focus the laser beam.

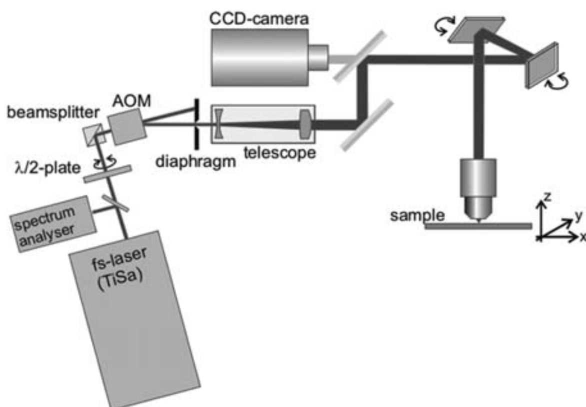


Fig. 12.8 Schematic representation of the experimental setup used for 2 PP photofabrication

12.3.3 Materials Used for 2PP

A great advantage for the application of the 2 PP technology in microstructure fabrication is provided by the wide variety of available materials, with properties that can fit almost any end application. By using fabricated structures as templates, it is also possible to fabricate 3D structures with high resolution from materials that cannot be directly patterned by the 2 PP technique. In this case the 3D structure is infiltrated with a required material, and in the subsequent step, the original structure is removed (chemically or thermally), resulting in an inversed 3D replica, which consists of a material used for the infiltration (see Section 12.4.1). Most of the photosensitive materials, which are currently used for the 2 PP microstructuring, were originally developed for lithography. Since virtually any photosensitive material can be structured by the 2 PP technique, there are still many unexplored materials. Figure 12.9a shows the subdivision of photosensitive materials used in our studies into different groups. Conventionally, photosensitive materials are subdivided into two classes: positive and negative resists. The illuminated volume of negative resists is cross-linked and unexposed material is removed during the sample development step (Fig. 12.9b). In positive resists, light induces dissociation of the molecules and the irradiated area is removed during the development step (Fig. 12.9b). Most positive resists are developed for the fabrication of integrated circuits by photolithography, where they are patterned in 2D and are used as a sacrificial layer in a lift-off process. Therefore, these resists are designed for an easy chemical or thermal removal. By applying femtosecond laser pulses one can write 3D structures in positive resists, in this case we are talking about two-photon activated processing, since no polymerization is actually taking place.

Negative resist materials can be roughly subdivided into solid and liquid, in accordance with the appearance of the material during the 2 PP processing. The solid materials presented in Fig. 12.9a are epoxy-based photoresists polymerized through cationic polymerization. Interaction with light generates an acid in the illuminated regions, the refractive index stays unchanged, and the appearance of patterns cannot be observed in real time. Actual polymerization takes

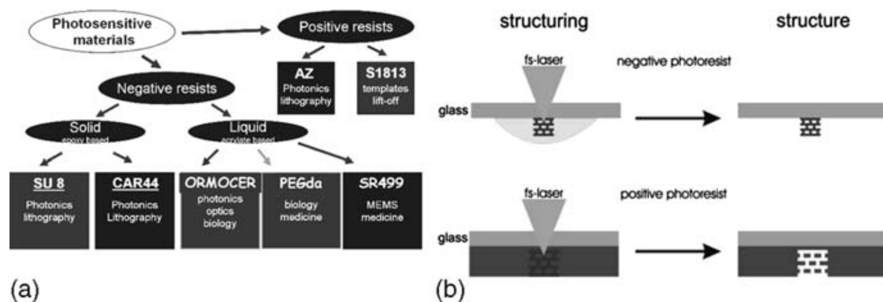


Fig. 12.9 (**a**) Materials for 2 PP technique; (**b**) negative and positive resist material processing

place during the post-bake processing step. After the post-bake processing step, the nonpolymerized material is removed by an appropriate developer.

Liquid materials presented in the diagram (Fig. 12.9a) are methacrylate (ORMOCER®) and acrylate (PEGda, SR499) based materials, which are polymerized via free-radical polymerization (see Section 12.2). These materials contain 1.8% of the photoinitiator 2-benzyl-2-dimethylamino-4′-morpholino-butyrophenone (Irgacure 369, Ciba SC, Switzerland), which is sensitive to the light of around and below 320 nm. ORMOCER (organically modified ceramics) designates a whole class of materials developed by "Fraunhofer Institute fuer Silicat Forschung" (Würzburg, Germany). Properties of these inorganic–organic hybrid materials can be tailored by means of chemical design fitting a wide variety of applications. We have investigated poly(ethylene) glycol diacrylate (PEGda, available under the name of SR610 from Sartomer Corporation, USA) for possible applications in biology and medicine. SR499 (Sartomer Corporation, USA) is used for the fabrication of micro-electromechanical systems (MEMS) and bio-MEMS.

12.4 Applications of 2 PP Technique

Despite the fact that 2 PP is a relatively new technology its application area has been rapidly expanding in the recent years. Fabrication of 3D photonic crystals by the 2 PP technique has been first proposed and demonstrated by Maruo et al. [10] and by now is applied by different groups in the world. Apart from that, 2 PP is also used for the fabrication of micromechanical systems [11], microfluidic devices [12], microoptical components [13], plasmonic components [14], biomedical devices [15], scaffolds for tissue engineering [16], and even natural proteins [17]. Microlasers have been demonstrated by 2 PP of optical gain medium by Yokoyama et al. [18]. In the following section, we will discuss in detail some applications of the 2 PP technique which have been studied by our group.

12.4.1 Applications in Photonics

One of the first and the most thoroughly studied applications of 2 PP technique is the fabrication of 3D photonic crystals. A photonic crystal is an artificial structure exhibiting periodic variation of the dielectric constant of material [19, 20]. Such structure has a similar effect on propagation of photons as the periodic variation of electric potential in regular crystals on propagating electrons, hence the name photonic crystal. As a result, a photonic bandgap – a frequency range, for which the propagation of photons in a certain direction is forbidden – occurs. The central position and the relative width of such a bandgap depend on the dielectric contrast and the periodicity of the structure.

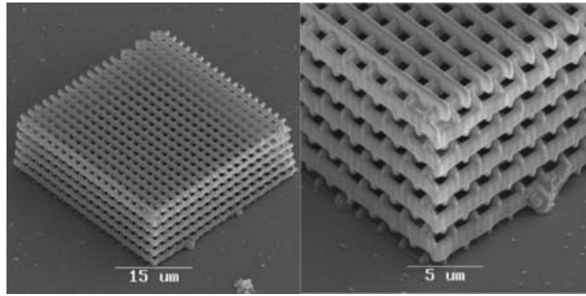


Fig. 12.10 Woodpile structure of hybrid organic–inorganic polymer fabricated by 2 PP exhibits periodic dielectric constant variation in any direction of propagation

An example of 1D photonic crystal is a dielectric mirror, which exhibits a periodic variation of the refractive index of the material in one direction. By changing the applied material and design one can fabricate mirrors that reflect light effectively in a wide range of wavelengths.

In a 3D photonic crystal, the refractive index changes periodically in any given direction. An example of such a structure is shown in Fig. 12.10. It is a so-called woodpile configuration with periodicity of the dielectric constant created by alternating polymer/air regions. The main distinction of a 3D photonic crystal is that one can design structures where photonic bandgaps for different light propagation directions overlap, resulting in a complete or omnidirectional photonic bandgap – a frequency range for which the propagation of light is forbidden in any direction. Devices based on photonic crystals allow tailoring propagation of light in a desired manner. Many fascinating physical phenomena occur in such structures: control of spontaneous emission [21], sharp bending of light [22], lossless guiding [23], zero-threshold lasing [24], trirefringence [25]. Futuristic prospects include not only applications in telecommunications as all-optical signal processing, but also "transistors" for light and optical computers.

It is worthwhile to mention that not all practical applications require omnidirectional bandgap. Function of many photonic devices can rely, for example, upon strong isotropy or low group velocity at the band edge. By utilizing these properties, collimators, dispersion compensators, multiplexers using superprism phenomena, photonic crystal-based optics, and other devices can be realized [26, 27, 28, 29, 30].

The central wavelength of a photonic bandgap for a given photonic crystal structure coincides approximately with its period. Therefore, in order to fabricate photonic crystals, performing in the visible or near-IR frequency range, structural resolution better than 1 μm is required. The ability of 2 PP to create complex volumetric structures with exceptionally high resolution makes this technology advantageous for the fabrication of 3D photonic crystals with bandgaps in the visible and near-infrared spectral ranges. It also implies that one is able to introduce defects at any desired locations, which is crucial for the practical applications of photonic crystals. The number of groups around the

world which apply 2 PP technique is increasing rapidly. Most of the photonic crystal configurations ever proposed by theoreticians have been realized [31, 32, 33, 34]; many of them would be impossible to produce by means of other technologies.

The experimental characterization of 3D photonic crystals is performed by means of the FTIR spectroscopy along a certain direction in a crystal. Due to the bandgap the spectra shall exhibit the dip in the transmission and an according peak in the reflection spectra. The results of the FTIR measurements on woodpile structures fabricated by the 2 PP technique are shown in Fig. 12.11. The rod distances were varied between 1.2 and 1.8 μm, the spectra indicate the clear bandgap positions with central frequency shifting to shorter wavelengths as the rod distance is reduced. This behaviour is in accordance with the theoretical predictions. In addition, the spectra show appearance of the higher order bandgaps in all samples, indicating the high quality of the fabricated structures. The absorption bands at around 3 and 3.4 μm come from the absorption of the material, as has been confirmed by measurements on the flat, unstructured layers.

Much effort has been made in order to fabricate 3D photonic crystals with the complete bandgaps located in the near-IR region, at the telecommunication wavelengths. The main challenge results from the fact that the structural resolution of such photonic crystals approaches the limit of the structural

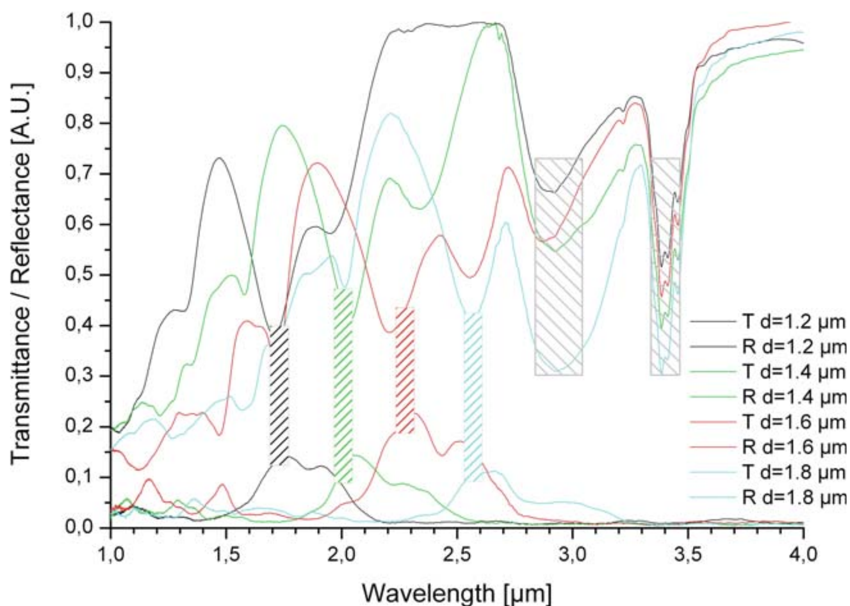


Fig. 12.11 Transmission and reflection spectra of woodpile photonic crystals with different rod spacings (*See* Color Insert)

resolution of the 2 PP technology and the applied materials. Further shift of the photonic bandgap central frequency into the visible range requires further downscaling of the structures and imposes an even bigger challenge on the current fabrication approaches and applied materials. Commonly used photosensitive materials exhibit shrinkage leading to a distortion of the fabricated structures which can result in a closing of a photonic bandgap. In order to avoid such distortions the structure can be pre-compensated [35] or mechanically stabilized by providing a massive frame around it [36].

The main drawback of 2 PP is a low refractive index contrast of the fabricated structures resulting from a low dielectric constant of the photosensitive materials. The relative width and position of a photonic bandgap depend on the refractive index contrast between two materials, in most cases between the air and the dielectric. It also imposes a restriction on the minimal refractive index value, required to obtain an omnidirectional bandgap. For the case of the woodpile topology, for example, material refractive index of at least 2.7 is required. The refractive indices of most photosensitive materials, which can be used for 2 PP applications, are far below this value.

One possibility to solve this problem is to use a 2 PP fabricated structures as templates for production of the inversed structures from high refractive index materials. Few groups have shown successful application of this approach using various photonic crystal structures as templates [37, 38, 39]. In order to fabricate a replica one has to be able to not only infiltrate the template with high refractive index material, but also remove the original structure. Due to their low chemical stability, positive resists are very attractive for this approach. As was described in Section 12.3.3 the illuminated area of positive resist materials is removed during the development processing step. Therefore, by writing a woodpile structure into such a material one obtains a direct hollow replica structure, as shown in Fig. 12.12a,b. After filling this structure with another material one can remove the original template by simply dissolving it in acetone or in a solution of NaOH. An example of a resulting replica in acrylate monomer is shown in Fig. 12.12c. This approach provides a prospect for fabrication

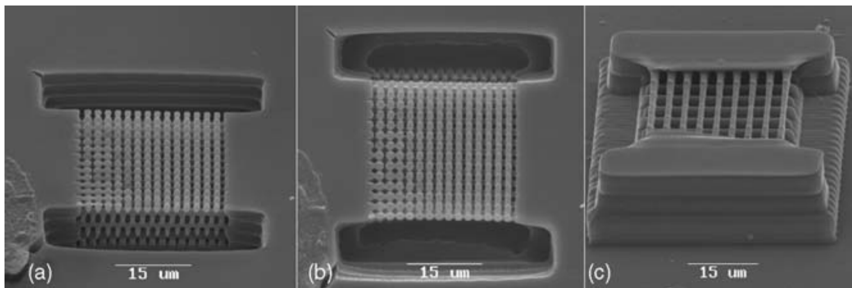


Fig. 12.12 Structuring of positive resist material: (**a**, **b**) woodpile structure in a positive resist; (**c**) replica in acrylate monomer (structure in **c** is replicated from a template different than **a**, **b**)

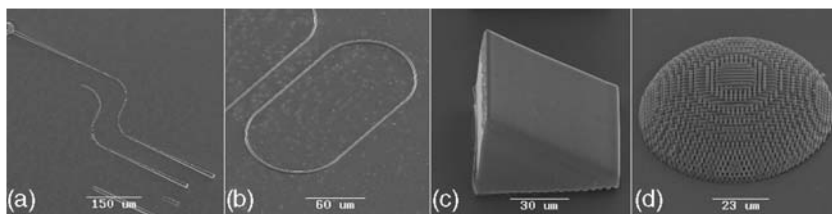


Fig. 12.13 Microoptical components fabricated by 2 PP technique: (**a**) tapered waveguides; (**b**) ring resonators: (**c**) microprism; (**d**) lens-shaped woodpile structure

of 3D structures with high resolution from materials that cannot be directly patterned by the 2 PP technique.

Due to their excellent optical properties, ORMOCERs are very attractive materials for the fabrication of microoptical devices. Using 2 PP, it is straightforward to generate complicated 2D and 3D structures, like ring resonators and 3D waveguide tapers, shown in Fig. 12.13a,b. These waveguides are suitable for guiding single-mode light at 1550 nm and can be used as on-a-chip interconnects. The main advantages of 2 PP are the high resolution of fabricated structures and the ability to produce 3D tapered waveguides, which can be used as mode converters. In addition, fabrication of diffractive and refractive microoptical elements is of great importance. Simple microprisms (Fig. 12.13c) and complex lens shaped woodpile (Fig. 12.13d) elements can also be rapidly fabricated with 2 PP technique. Using 2 PP one can create large-area complex design 2D microoptical element arrays that can then be used as a master for microimprinting or injection-moulding replication in mass production.

12.4.2 Biomedical Applications

We have demonstrated several very promising biomedical applications of 2 PP technique: for tissue engineering, drug delivery, and medical implants. Artificial fabrication of a living tissue that will be able to integrate with the host tissue inside a body is a bold and challenging task undertaken by tissue engineering. Natural repair of a tissue at the particular site is a result of complex biological processes, which are currently the subject of intensive research and are not yet fully understood. In order to encourage cells to form tissue, one has to create an appropriate environment, exactly resembling that of a particular tissue type. Some cell types can preserve tissue-specific features in a 2D environment, while others require a 3D environment. One of the most popular approaches in tissue engineering is the use of 3D scaffolds whose function is to guide and support cell proliferation in 3D. The ability to produce arbitrary 3D scaffolds is therefore very appealing. Few techniques that can create 3D porous scaffolds have been developed in the recent years [40, 41, 42, 43, 44, 45]. These techniques can be

subdivided into passive and active. The passive techniques, such as phase separation, yield porous structures with high resolution and uniform pore size. However, they do not allow fabrication of exactly identical structures and provide little control over the location of individual pores. On the other hand, the active techniques, such as inkjet printing and stereolithography, provide possibility to produce any CAD designed structure, but have resolution of the order of tens of micrometres. Advantage of 2 PP for the fabrication of scaffolds is a combination of unprecedented resolution, high reproducibility confidence, and a possibility to fabricate true 3D structures. Therefore, scaffolds fabricated by the 2 PP technique will enable systematic studies of cell proliferation, acquired functionality, and tissue formation in 3D. Figure 12.14a,b show an original CAD design and a scanning electron microscope (SEM) image of a fabricated structure, which resembles pulmonary alveoli – microcapillaries responsible for gas exchange in the mammalian lungs. A 3D polymeric mesh structure fabricated from ORMOCER resembles the interconnected pores that are found in the bones (Fig. 12.14c,d).

2 PP technique can also be applied for the fabrication of implants and prostheses. For example, the malleus, incus, and stapes bones serve to transmit sounds from the tympanic membrane to the inner ear. Ear diseases may cause discontinuity or fixation of the ossicles, which results in conductive hearing loss. The size of the total ossicular replacement prosthesis (TORP) is of the order of few millimetres and varies from patient to patient. The materials used in ossicular replacement prostheses must demonstrate appropriate biological compatibility, acoustic transmission, stability, and stiffness properties. The prostheses prepared using Teflon®, titanium, Ceravital, and other conventional materials have demonstrated several problems during clinical studies, including migration, puncture of the eardrum, difficulty in shaping the prostheses, and reactivity with the surrounding tissues.

We have demonstrated the application of the 2 PP technique for rapid prototyping of ORMOCER middle-ear bone replacement prostheses [46]. Figure 12.15 shows an original CAD design and an optical microscope image of the fabricated TORP. The 2 PP technique provides several advantages over

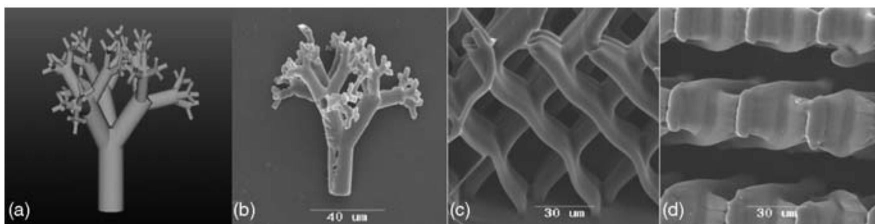


Fig. 12.14 Microstructures for tissue engineering fabricated by 2 PP technique: (**a**) original CAD design of microcapillaries; (**b**) corresponding structure fabricated by 2 PP; (**c, d**) different orientation views of 3D scaffolds with interconnected pores

Fig. 12.15 Middle-ear bone replacement prosthesis: (**a**) original CAD design; (**b**) optical microscope image of a structure fabricated by 2 PP in ORMOCER; (**c**) in vitro implantation of the fabricated implant

the conventional processing for scalable mass production of ossicular replacement prostheses. First, the raw materials used in this process are widely available and inexpensive. Second, 2 PP can be set up in a conventional clinical environment (e.g. an operating room) that does not contain clean room facilities. It allows fabrication of patient-specific prosthesis based on the optical coherence tomography (OCT) analysis or other provided data. Moreover, the resolution required for TORPs is not very high and therefore one can accelerate the fabrication process even further. And finally, 2 PP of ossicular replacement prostheses is a straightforward, single-step process, as opposed to the conventional multistep fabrication techniques. We anticipate that the number of applications of the 2 PP technique for the fabrication of prosthesis will rapidly increase in the nearest future.

Transdermal drug delivery is a method that has experienced a rapid development in the past two decades and has often shown improved efficiency over the other delivery routes [47]. It avoids many issues associated with intravenous drug administration, including pain to the patient, trauma at the injection site, and difficulty in providing sustained release of pharmacologic agents. In addition, precise dosing, safety, and convenience are also addressed by transdermal drug delivery. However, only a small number of pharmacological substances are delivered in this manner today. The most commonly known example is nicotine patches. The main reason for that is the significant barrier to diffusion of substances with higher molecular weight provided by the upper layers of the skin. The top layer, called stratum corneum, is composed of dead cells surrounded by lipid. This layer provides the most significant barrier to diffusion to approximately 90% of transdermal drug applications [48, 49]. A few techniques enhancing the substance delivery through the skin have been proposed. Two of the better-known active technologies are iontophoresis and sonophoresis. The rate of product development involving these technologies has been relatively slow [50, 51]. This is partly conditioned by the relative complexity of the

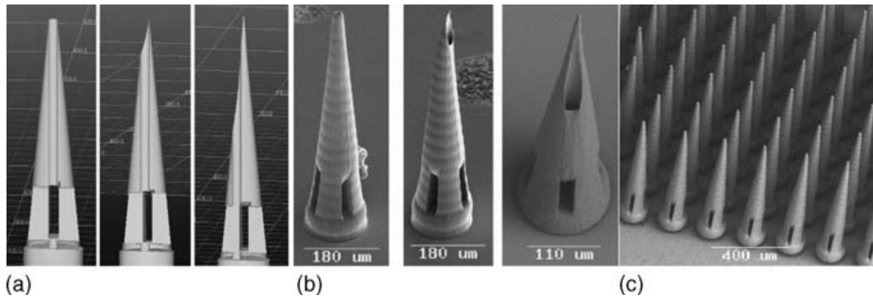


Fig. 12.16 Hollow microneedles for transdermal drug delivery: (**a**) cross-section of original CAD design of microneedles with different channel positions and tip sharpnesses; (**b**) SEM images of respective microneedles fabricated by 2 PP technique; (**c**) an array of microneedles fabricated by 2 PP technique

resulting systems, compared to the passive transdermal systems. One of the passive technologies is based on microneedle-enhanced drug delivery. These systems use arrays of hollow or solid microneedles to open pores in the upper layer of the skin and assist drug transportation. The length of the needles is chosen such that they do not penetrate into the dermis, pervaded with nerve endings, and thus do not cause pain. In order to penetrate the stratum corneum, microneedles for drug delivery have to be longer than 100 μm and are generally 300–400 μm long, since the skin exhibits thickness values that vary with age, location, and skin condition. Application of microneedles has been reported to greatly enhance (up to 100,000 fold) the permeation of macromolecules through the skin [52]. The microneedles for withdrawal of blood must exceed lengths of 700–900 μm in order to penetrate the dermis, which contains blood vessels. Most importantly, microneedle devices must not fracture during penetration, use, or removal.

The flexibility and high resolution of the 2 PP technique allow rapid fabrication of microneedle arrays with various geometries (Fig. 12.16) and to study its effect on the tissue penetration properties. Results of our studies indicate that microneedles created using the 2 PP technique are suitable for in vivo use and for integration with the next generation MEMS- and NEMS-based drug delivery devices.

12.5 Summary and Outlook

An essential factor for the progress in the nanotechnology and its driving power is the development of high-fidelity nano- and micro-fabrication techniques. Femtosecond laser technologies based on nonlinear light–matter interactions provide possibility of cost-efficient manufacturing with high resolution and unprecedented flexibility. Increased attention to these technologies from the

industry representatives is stimulated by the recent development of compact turnkey ultrashort pulse laser systems.

In contrast to the techniques based on material ablation, two-photon polymerization (2 PP) allows the fabrication of true 3D structures with a resolution down to 100 nm. The versatility of the 2 PP technology and the large number of applicable materials contribute to the wide range of applications of this technology which are rapidly growing.

Acknowledgments The authors gratefully acknowledge very important contribution from their colleagues, who have been involved in different parts of this work: J. Serbin, C. Reinhardt, S. Passinger, R. Kiyan, and R. Cotton. Biomedical applications of the 2 PP technique have been studied in cooperation with A. Doraiswamy, T. Platz, R. Narayan, R. Modi, R. Auyeung, D.B. Chrisey, and O. Adunka. This work has been supported by the DFG "Photonic crystals" research program SPP1113.

References

1. KM Labs
2. "Leistungsbilanz der allgemeinen Stromversorgung in Deutschland 2004 bis 2010" Verband der Netzbetreiber – VDN – e.V. beim VDEW (2003)
3. Zewail AH (2000) *Angew. Chem. Int.* Ed., 39:2587–2631
4. Göppert-Mayer M (1931) *Ann. Physik*, 9:273
5. Kaiser W, Garrett CGB (1961) *Phys. Rev. Lett.* 7:229
6. Shen YR (1984) The principles of nonlinear optics. Wiley, New York
7. Sun HB, Kawata S (2004) Two-photon polymerization and 3D lithographic Microfabrication, In: N. Fatkullin (ed.), *NMR, 3D Analysis, Photopolymerization*, Springer, pp 169–273
8. Ovsianikov A, Passinger S, Houbertz R, Chichkov BN (2006) Three-dimensional material processing with femtosecond lasers, In: Claude R. Phipps (ed.), *Laser Ablation and its Applications*, Springer Series in Optical Sciences, pp 129–167
9. Bertsch A, et al (2002) *MRS Symp. Proc.* 758
10. Maruo S, Nakamura O, Kawata S (1997) *Opt. Lett.* 22:132
11. Sun HB et al. (2000), *Opt Lett* 25:1110
12. Maruo et al. (2003) *MRS Symp. Proc.* 739
13. Gu M et al. (2006) *Optics Express* 14(2):810
14. Reinhardt C et al. (2006) *Optics Lett.* 31:1307
15. Doraiswamy et al. (2006) *Acta Biomaterialia* 2:267–275
16. Schlie S et al. (2007) Three-dimensional cell growth on structures fabricated from ORMOCER® by two-photon polymerization technique, J Biomater Appl 22: 275–287
17. Basu et al. (2004) *Biomacromolecules* 5(6)
18. Yokoyama et al. (2003) *Appl. Phys. Lett.* 82(19)
19. Yablonovitch E (1987) *Phys. Rev. Lett.* 58:2059
20. John S (1987) *Phys. Rev. Lett.* 58:2486
21. Joannopoulos JD, Villeneuve PR, Fan S (1997) Photonic crystals: putting a new twist on light. Nature 386:143–149.
22. Noda S et al. (2000) Science 289:604
23. Chutinan A, John S, Toader O (2003) *Phys. Rev. Lett.* 90:123901
24. Markowicz P et al. (2002) *Opt. Lett.* 27: 5, 351
25. Netti MC et al. (2001) *Phys. Rev. Lett.* 86:1526

26. Kosaka et al. (1999) *Appl. Phys. Lett.* 74:1212
27. Joannopoulos J (1995) Photonic crystals: molding the flow of light. Princeton University Press
28. Johnson S, Joannopoulos J (2005) Photonic crystals: the road from theory to practice. Springer, New York
29. Busch K (2004) Photonic crystals: advances in design, fabrication, and characterization, Wiley-VCH
30. Noda S (2003) Roadmap on photonic crystals, Kluwer Academic Publishers
31. Kaneko et al. (2003) *Appl. Phys. Lett.* 83(11)
32. Deubel et al. (2004) *Appl. Phys. Lett.* 85(11)
33. Serbin J et al. (2004) *Optics Express* 12(21):5221
34. Seet et al. (2006) *Appl. Phys. A.* 82:683
35. Sun H-B et al. (2004) *Appl. Phys. Lett.* 85:17, 3709
36. Deubel M et al. (2004) *Nat. Mater.* 3:444
37. King JS et al. (2005) *Appl. Surface Sci.* 244:511–516
38. Emelchenko et al. (2005) *J. Opt. A: Pure Appl. Opt.* 7:213
39. Tetreault et al. (2006) *Adv. Mater.* 18:457–460
40. Sachlos E, Czernuszka JT (2003) *Eur. Cells and Mater.* 5:29
41. Prendergast PJ, McHugh PE (2004) *Topics in bio-mechanical engineering*, pp 147–166. Trinity Centre for Bioengineering & National Centre for Biomedical Engineering Science
42. Tsang VL, Bhatia SN (2006) "Fabrication of Three-Dimensional Tissues, *Adv. Biochem. Engin./Biotechnol.* 103:189
43. Vozzi G, Ahluwalia A (2007) *J. Mater. Chem.* 17:1248
44. Hull C (1990) Method for production of three-dimensional objects by stereolithography, US Patent 4929402
45. Hutmacher et al. (2004) *Trends in Biotechnol.* 22(7)
46. Ovsianikov et al. (2007) Proceedings of 5th International Conference on Photo-Excited Processes and Applications, to be published in "Applied Surface Science"
47. Chong S, Fung HL (1989) Transdermal drug delivery systems: pharmacokinetics, clinical efficacy, and tolerance development. In: J Hadgraft, RH Guy (eds.), *Transdermal Drug Delivery: Developmental Issues and Research Initiatives*. Marcel Dekker, New York, p 135
48. Flynn GL (1996) Cutaneous and transdermal delivery: Processes and systems of delivery. In: GS Banker, CT Rhodes (eds.), *Modern Pharmaceutics*. Marcel Dekker, New York, pp 239–299
49. Sivamani RK et al. (2005) *Skin Res. Technol.* 11(2):152–156
50. Mitragotri S (2001) *J. Controlled Rel.* 71:23–29
51. Guy RH (1998) *J. Pharm. Pharmacol.* 50(4):371–374
52. Barry BW (2001) *Eur. J. Pharm. Sci.* 14:101–114

Index

A

Absorption spectra, semiconductor
  coherent and incoherent carrier correlations, 322–323
  excitation-induced dephasing, 326–328
  excitonic states, 319–320
  linear optical polarization, 323–326
  pump-probe calculations, 320–322
  semiconductor Bloch equations (SBE), 316–319
Acoustic microscopy, 234–235
Acousto-optic modulator (AOM), 435
Atomic force microscopy (AFM), 230, 239
  force-displacement curves and, 230–232
  lateral force imaging and, 232–233
  modulated nanoindentation and, 233–234

B

Band alignment at, 183
  HfO_2/Mo interface, 187–188
  Si/HfO_2 interface, 184–186
  SiO_2/HfO_2 interface, 186
Band-edge lasers, 391–392
Band-to-band recombination, 192
Bernoulli–Euler theory, 236–237
Biot–Savart law, 93
Bloch theorem, 261, 356, 368
Boltzmann transport equation, 123
  semi-classical, 125–129
Boolean logic circuits, 37
Born–Oppenheimer approximation, 173, 257
Bose–Einstein distribution, 283
Brillouin zone, 358
  photonic crystal slab, 362
BTE, *see* Boltzmann transport equation

C

Carbon nanotubes, transport in, 140–141
Cauchy's integral theorem, 301
CB, *see* Coulomb blockade
CCVT method, *see* Constant capacitance voltage transient method
CCWs, *see* Coupled-cavity photonic crystal waveguides
Chalcogenide glasses, 19
Classical and quantum optics, of semiconductor nanostructures, 255
  absorption spectra
    coherent and incoherent carrier correlations, 322–323
    excitation-induced dephasing, 326–328
    excitonic states, 319–320
    linear optical polarization, 323–326
    pump-probe calculations, 320–322
    semiconductor Bloch equations (SBE), 316–319
  carriers and exciton populations, radiative recombination of, 333–335
  incoherent populations, probing, 337
    exciton correlations, dynamics of, 338–339
    linear terahertz response, 342–347
    terahertz spectroscopy of excitons, 340–342
  interactions
    complete system Hamiltonian in different dimensions, 302–304
    Coulomb interaction, 299–302
    electric dipole interaction, 294–296
    light–matter interaction, 285–294
    many-body Hamiltonian, 283–285
    phonon–carrier interaction, 296–299

Classical and quantum optics (*cont.*)
luminescence equations, 329–333
quantization, of quasi-particles, 256
of electromagnetic fields, 273–278
electron density of states, 271–272
electrons in periodic lattice potential, 260–261
k p perturbation theory, 261–264
second, of carrier system, 265–266
second, of lattice vibrations, 278–283
system Hamiltonian in first, 258–260
systems with reduced effective dimensionality, 267–271
quantum dynamics, 304–305
cluster expansion, 310–314
commutator properties, 305–307
general operator dynamics, 307–310
singlet–doublet correlations, 314–315
quantum-well emission, correlated photons in, 335–337
Cluster-expansion solution, quantum dynamics and, 304–305, 310–314
commutator properties, 305–307
general operator dynamics, 307–310
singlet–doublet correlations, 314–315
CMOL circuits, 15, 24
CMOL DSP circuits, 47–52
CMOL FPGA circuits, 35–46
and HP's FPNI circuits, 26–27
memory circuits, 28–35
CMOS-to-nanowire interfaces, 23–27
CNTs, *see* Carbon nanotubes
Complete system Hamiltonian, in different dimensions, 302–304
Constant capacitance voltage transient method, 200
Coulomb blockade, 157–160, 161–162, 164, 196, 210
Coulomb interaction, 299–302
Coupled-cavity photonic crystal waveguides, 398
Coupling amplitudes, of excited photonic crystal modes, 373–374

D
Datta–Das spin transistor, 94
Deep level transient spectroscopy, 199
Deep ultraviolet (DUV) photolithography, 386–387
Density functional theory (DFT), of high-*k* dielectric gate stacks, 171
ab initio packages, 179–180
band alignment at, 183
calculation, 178–179
HfO_2/Mo interface, 187–188
Si/HfO_2 interface, 184–186
SiO_2/HfO_2 interface, 186
electrons and phonons, 172–173
energy minimization and molecular dynamics, 176
many-electron problem and, 173–175
pseudopotential, 175
recent theoretical results, 180–183
supercell/slab technique, 176–178
Digital electronics, semiconductor/nanodevice hybrids circuits for, 15
CMOL DSP circuits, 47–52
CMOL FPGA circuits, 35–46
CMOL memories, 28–35
defect tolerance, 21–22
devices for, 17–21
micro-to-nano interface, 23–27
regularity, 22–23
Dilute magnetic semiconductor (DMS), 87–88
Dirac equation, 62
DLTS, *see* Deep level transient spectroscopy
Dresselhaus spin–orbit field, 93

E
ECC techniques, 29
Electric dipole interaction, 294–296
Electromagnetic fields, quantization, 273–278
Electron-beam lithography, 380–381
Electronic transport, in nanoscale systems, 122–125
Electron microscopy
mechanical resonance, 236–237
in situ tensile/bending test, 237–238
Electro-optics (EO), photonic crystals and, 412–414
Elliott formula, 321
Envelope-function approximation, 267–270
Excitation-induced dephasing, 326–328
see also Absorption spectra, semiconductor

F
Fermi's Golden rule, 126
Ferromagnetic proximity polarization, spin extraction and, 97–99
Ferromagnetic semiconductors, 83–84
spin extraction and ferromagnetic proximity polarization, 97–99
spin injection, 96–97

Field-programmable gate arrays (FPGA), 35
Finite-difference time domain (FDTD) method, 364
Fowler-Nordheim tunneling, 207–208
FPP, *see* Ferromagnetic proximity polarization

G
Giant magnetoresistance (GMR), 68–70
Goeppert Mayer gauge transformation, 294
Goto-pair operation, 37

H
Hanle effect, 102–104
Hartree–Fock factorization, 311
Heisenberg uncertainty principle, 62, 275
Hellman–Feynman theorem, 176
Helmholtz equation, 274
HfO_2/Mo interface, band alignment at, 187–188
High-*k* dielectric gate stacks, DFT of, 171
  ab initio packages, 179–180
  band alignment at, 183
    calculation, 178–179
    HfO_2/Mo interface, 187–188
    Si/HfO_2 interface, 184–186
    SiO_2/HfO_2 interface, 186
  electrons and phonons, 172–173
  energy minimization and molecular dynamics, 176
  many-electron problem and, 173–175
  pseudopotential, 175
  recent theoretical results, 180–183
  supercell/slab technique, 176–178
High-resolution transmission electron microscope (HRTEM), 140
High-speed, low-voltage silicon photonic crystal modulator, 400–404
Hohenberg-Kohn theorem, 173-174
Holography, 381–382
Hooke's law, 233, 239
HP's FPNI circuits, 26–27

I
Interlayer exchange coupling (IEC), 65–68

J
"Johnson–Silsbee" geometry, 101–102
  see also Lateral spin transport devices
Julliere model, for TMR, 73–74

K
Kerr lens effect, 434
Kohn–Sham (KS) equations, 174, 176
k p perturbation theory, 261–264
Kramer's degeneracy, 60

L
Landauer–Büttiker formula, 151, 153
Landau-Lifshitz-Gilbert (LLG) magnetization dynamics, 79–80
Laser direct writing, by two-photon absorption, 382–383
  see also Photonic crystals
Lateral spin transport devices, 99–100
  Hanle effect, 102–104
  lateral spin valve, 100
  non-local geometry, 101–102
  spin Hall effect, 104–108
  see also Spintronics
LDA, *see* Local density approximation
Light-emitting diodes (LEDs), photonic crystals and, 392–393
Light–matter interaction, 285–294
Linear muffin-tin orbitals in atomic sphere approximation (LMTO-ASA) method, 180–181
Linear optical polarization, 323–326
  see also Absorption spectra, semiconductor
Linear terahertz (THz) response, 342–347
Linear waveguides, in photonic crystal slabs, 362–364
Liouville–von Neumann equation, 122–23, 126
Local density approximation, 174
Low-energy electron point source (LEEPS) microscope, 214

M
Magnetic hard drives, 82–83
Magnetic Random Access Memory, 40–86
Magnetic tunnel junction, 71–73
  Julliere model for TMR, 73–74
  MgO-based, 74–75
Magneto-optic Kerr effect, 67
Many-body Hamiltonian, 283–285
Maxwell-semiconductor Bloch equations, 319
MBE, *see* Molecular beam epitaxy
Metal-induced gap state (MIGS) model, 183
Metal-organic chemical vapor deposition (MOCVD), 388

Methacrylate (ORMOCER®), 437
MgO-based MTJ, 74–75
Microcavities, in photonic crystal slabs, 364–365
Millipede concept, 242
MOKE, *see* Magneto-optic Kerr effect
Molecular beam epitaxy, 63, 66, 133
Moore Law, 16
MRAM, *see* Magnetic Random Access Memory
MSBE, *see* Maxwell-semiconductor Bloch equations
MTJ, *see* Magnetic tunnel junction
Multi-photon absorption (MPA), 428

N
Nanocrystalline semiconductors, trapping phenomena in, 191–197
  applications, 205–219
  classical investigation methods, 198
    constant capacitance voltage transient (CCVT) method, 200
    deep level transient spectroscopy (DLTS), 199
    photoinduced current transient spectroscopy (PICTS), 200
    thermally stimulated currents (TSC) method, 201
    thermally stimulated depolarization currents (TSDC) method, 202
Nanocrystalline silicon, 205–206, 211
Nano-electromechanical system
  aplications
    catalysis, 246
    electrical power generation, 247–248
    nanolithography and high-density data storage, 241–242
    nanomanipulators, 244–246
    optics and telecommunications, 243–244
    sensors, 239–241
  electron microscopy, 236–238
  nanoscale, mechanical properties at, 223–225
    defects, 227–229
    phase transitions, 229
    surface effects, 225–226
  optical methods, 238–239
  scanning probe-based methods
    contact stiffness mapping, 234–235
    force-displacement curves, 230–232
    lateral force imaging, 232–233
    modulated nanoindentation, 233–234
Nanoimprint lithography, 384–385
Nanomanipulators, NEMS and, 244–246
Nanoscale systems, transport in, 115
  electronic transport and, 122–125
  non-classical and quantum effect devices, 119–122
  quantum confined systems, diffusive transport in
    effects of, 129–143
    semi-classical Boltzmann transport equation, 125–129
  semiconductor device scaling, 117–119
  single electron tunneling
    Coulomb blockade, 157–160, 161–162, 164
    SET modeling and simulation, 160–161
    Si nanoelectronic devices, 162–163
    single electron phenomena, 154–156
  transmission and, 142
    quantized conductance, 147–151
    quantum waveguides, 151–154
    vertical transport through heterostructures, 143–147
National Nanotechnology Institute (NNI), 3–4
nc-Si, *see* Nanocrystalline silicon
Negative refraction, 415–418
  see also Photonic crystals
NEMS, *see* Nano-electromechanical system
Neuromorphic computing, 11

O
One pin-nanowire-nanodevice-nanowire-pin link, 42–43
Optical charging spectroscopy (OCS), 203–205, 211–212
Optical coherence tomography (OCT) analysis, 443
Optics and telecommunications, NEMS and, 243–244

P
Passive photonic crystal waveguide Mach–Zehnder interferometers, 398–399
PBG, *see* Photonic band gap
PCW modulator, *see* Photonic crystal waveguide modulator
PDMS, *see* Polydimethylsiloxane
Phase prism, 414–415
Phonon–carrier interaction, 296–299
Phonon-operator dynamics, 308–309

Photoinduced current transient spectroscopy, 200
Photonic band gap
  physical origin, 355–358
  surface states in, 366–369
Photonic crystals, 353
  application
    in electro-optics, nonlinear optics, and sensing, 412–414
    to extraction efficiency of LEDs and VCSELs, 392–393
  band gap, physical origin, 355–358
  basic surface eigenmode equations, bulk eigenmode equations and, 368–371
  beam of finite width, transmission and refraction of, 407–410
  boundary equations, 373–374
  defect modes, control of light with, 359–360
  fabrication of
    electron-beam lithography, 380–381
    holography, 381–382
    laser direct writing by two-photon absorption, 383
    nanoimprint lithography, 384–385
    process integration, 386–387
    self-assembly and templating, 383–384
  filters, 393–396
  forward and backward eigenmodes, equal partition of, 369–371
  lasers
    band-edge, 391–392
    electrically pumped, 389–391
    optically pumped with cavity, 388–389
  modulators, 396–405
  negative refraction, 415–418
  phase prism, 414–415
  photonic band gap (PBG), surface states in, 366–367
  slabs
    band structures of, 361–362
    linear waveguides in, 362–364
    microcavities in, 364–365
    transmission theory for, 377–380
  superprism effect, 406–407
  surface orientation, mode degeneracy on, 371–373
  surface refraction/coupling, 374–376
  three-dimensional, 360–361
  two-dimensional, 358
  wavelength division multiplexing (WDM), 410–412
Photonic crystal waveguide modulator, 399–402
Photonics, 2 PP technique and, 439–444
PICTS, *see* Photoinduced current transient spectroscopy
Plane-wave expansion method, 364
Polydimethylsiloxane, 385
Poly (*N*-isopropylacrylamide) (PNIPAM), 245–246
2 PP, *see* Two-photon photopolymerization

Q
Quantum confined systems, diffusive transport in
  effects, 129–131
    quasi-1D systems, 136–140
    quasi-2D systems, 132–136
  semi-classical BTE, 125–129
Quantum-well emission, correlated photons in, 335–337
Quasi-2D systems, 132–136
Quasi-1D systems, transport in, 136
  carbon nanotubes (CNTs), 140–142
  self-assembled semiconductor nanowires, 138–139
  Si nanowires, 136–140
Quasi-particles quantization, in semiconductors, 256–258
  of electromagnetic fields, 273–278
  electron density of states, 271–272
  electrons in periodic lattice potential, 260–261
  k p perturbation theory, 261–264
  second
    of carrier system, 265–266
    of lattice vibrations, 278–283
  system Hamiltonian in first, 258–260
  systems with reduced effective dimensionality, 267–271

R
Rashba spin–orbit coupling, 93–95
Resonant tunneling diodes (RTDs), 136, 145–147
RKKY theory, *see* Interlayer exchange coupling (IEC)

S
SBE, *see* Semiconductor Bloch equations
Scanning probe-based nanolithography, 241–243

Schrödinger equation, 259
SDR, *see* Spin-dependent recombination
Self-assembled semiconductor nanowires, transport in, 140
Semiconductor Bloch equations, 316–319
Semiconductor luminescence equations, 329–333
Semiconductor/nanodevice hybrids circuits, for digital electronics, 15
CMOL DSP circuits, 47–52
CMOL FPGA circuits, 35–46
CMOL memories, 28–35
defect tolerance, 21–22
devices for, 17–21
micro-to-nano interface, 23–25
regularity, 22–23
Semiconductor nanostructures, classical and quantum optics, 255
absorption spectra
coherent and incoherent carrier correlations, 322–323
excitation-induced dephasing, 326–328
excitonic states, 319–320
linear optical polarization, 323–324
pump-probe calculations, 320–322
semiconductor Bloch equations (SBE), 316–319
carriers and exciton populations, radiative recombination of, 333–336
incoherent populations, probing, 337
exciton correlations, dynamics of, 338–339
linear terahertz response, 342–347
terahertz spectroscopy of excitons, 340–342
interactions in
complete system Hamiltonian in different dimensions, 302–304
Coulomb interaction, 299–302
electric dipole interaction, 294–296
light–matter interaction, 285–294
many-body Hamiltonian, 283–285
phonon–carrier interaction, 296–299
luminescence equations, 329–333
quantization, of quasi-particles in, 256
of electromagnetic fields, 273–278
electron density of states, 271–272
electrons in periodic lattice potential, 260–261
k p perturbation theory, 261–264
second, of carrier system, 265–266
second, of lattice vibrations, 278–283
system Hamiltonian in first, 258–260
systems with reduced effective dimensionality, 267–271
quantum dynamics, 304
cluster expansion, 310–314
commutator properties, 305–307
general operator dynamics, 307–310
singlet–doublet correlations, 314–315
quantum-well emission, correlated photons in, 335–337
Semiconductor spintronics, 86
ferromagnetic, 87–88
spin extraction and ferromagnetic proximity polarization, 93–95
spin injection, 97–98
spin coherence, optical studies of, 88–92
spin–orbit coupling, role of, 93–95
SEMPA, *see* Spinpolarized scanning electron microscopy with polarization analysis
SETs, *see* Single electron transistors
Si/HfO_2 interface, band alignment at, 184–186
Silicon CMOS circuits, 6
Silicon-on-insulator (SOI) MOSFETs, 16, 120, 163
Si nanowires (SiNW), transport in, 136–138
Single electron transistors, 136, 155, 210
Coulomb oscillations and, 158–160
modeling and simulation, 160–161
Single electron tunneling
Coulomb blockade, 157–160, 161–162
SET modeling and simulation, 160–161
Si nanoelectronic devices, 162–163
single electron phenomena, 154–156
SiO_2/HfO_2 interface, band alignment at, 186
SLE, *see* Semiconductor luminescence equations
Snell's law, 414–415
Spin coherence, optical studies, 88–92
spin–orbit coupling, role of, 93–95
Spin-dependent light-emitting diode (spin-LED), 96
Spin-dependent recombination, 214
Spin Hall effect, 104–108
Spin–orbit coupling, 93–95
Spinpolarized scanning electron microscopy with polarization analysis, 67
Spin torque, 75–77
origin, 77–79
precessional magnetization dynamics, excitation of, 79–81

Spintronics, 59–63
  lateral spin transport devices, 99–100
    Hanle effect, 102–104
    lateral spin valve, 100
    non-local geometry, 101–102
    spin Hall effect, 104–108
  metallic magnetic multilayers, 64
    applications, 81–86
    giant magnetoresistance (GMR), 68–70
    interlayer exchange coupling (IEC), 65–68
    magnetic tunnel junction (MTJ), 71–75
    spin torque, 75–81
    spin valves, 70–71
  semiconductor, 86
    ferromagnetic, 87–88
    ferromagnet/semiconductor structures, 96–99
    spin coherence, optical studies of, 88–95
Spin valve, 70–71
Stereolithography, 432
Supercell/slab technique, 176–178
Superprism effect, the, 406–407

T
TEM, *see* Transmission electron microscopy
Terahertz (THz) spectroscopy, of excitons, 340–342
Thermally stimulated depolarization currents method, 202
Thermo-optic photonic crystal waveguide modulators, 405
Three-dimensional photonic crystals, 360–361
Three-terminal ballistic junction (TBJ), 152–154
Time-of-flight (TOF) technique, 216
Time-resolved Faraday rotation, 89
TMR, *see* Tunneling magnetoresistance
Total ossicular replacement prosthesis (TORP), 443
Transmission electron microscopy, 227, 236
Transmission theory, for photonic crystal slab, 377–380
TRFR, *see* Time-resolved Faraday rotation
TSDC method, *see* Thermally stimulated depolarization currents method
Tsu-Esakt formula, 144–145
Tunneling magnetoresistance, 72–75
Two-dimensional photonic crystals, 358
Two-photon absorption (TPA), 428, 432
Two-photon photopolymerization (2 PP), 383, 430–433
  applications
    biomedical, 441–444
    in photonics, 437–441
  experimental setup for, 435
  materials used for, 436–437

U
Ultrashort laser pulses, generation, 433–434
Ultrasonic force microscope (UFM), 235

V
Vapor–liquid–solid (VLS) nanowires (NW), 138–139
Vertical cavity surface-emitting laser (VCSEL), photonic crystals and, 389, 393–396

W
Wannier equation, 319
Wavelength division multiplexing (WDM), 410–412

Nanostructure Science and Technology

(Continued from p. ii)

Current volumes in this series:

Nanoparticles: Building Blocks for Nanotechnology
Edited by Vincent Rotello

Nanostructured Catalysts
Edited by Susannah L. Scott, Cathleen M. Crudden, and Christopher W. Jones

Self-Assembled Nanostructures
Jin Z. Zhang, Zhong-lin Wang, Jun Liu, Shaowei Chen, and Gang-yu Liu

Polyoxometalate Chemistry for Nano-Composite Design
Edited by Toshihiro Yamase and M.T. Pope

Computational Methods for Nanoscale Applications: Particles, Plasmons and Waves
Igor Tsukerman

Nanoelectronics and Photonics: From Atoms to Materials, Devices, and Architectures
Edited by Anatoli Korkin and Federico Rosei

Color Insert

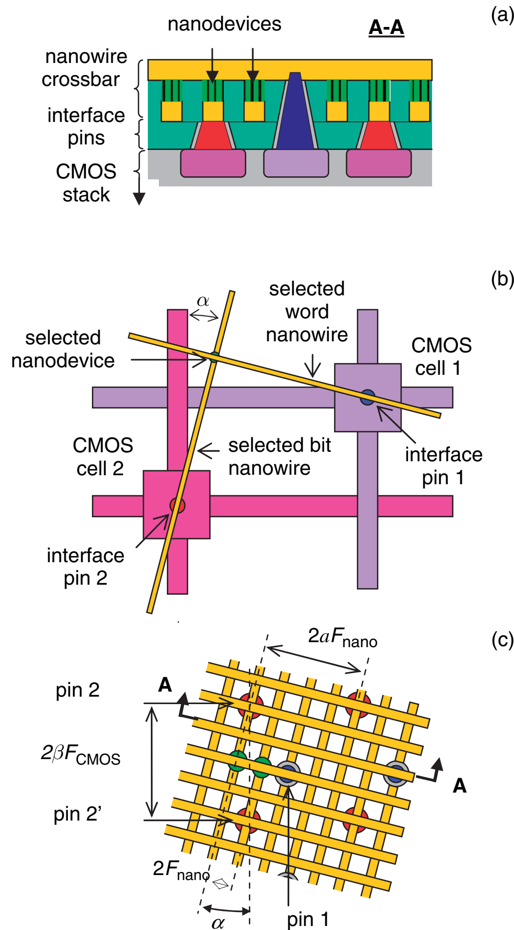


Fig. 4.4 The generic CMOL circuit: (**a**) a schematic side view, (**b**) a schematic top view showing the idea of addressing a particular nanodevice via a pair of CMOS cells and interface pins, and (**c**) a zoom-in top view on the circuit near several adjacent interface pins. On panel (**b**), only the activated CMOS lines and nanowires are shown, while panel (**c**) shows only two devices. (In reality, similar nanodevices are formed at all nanowire crosspoints.) Also disguised on panel (**c**) are CMOS cells and wiring

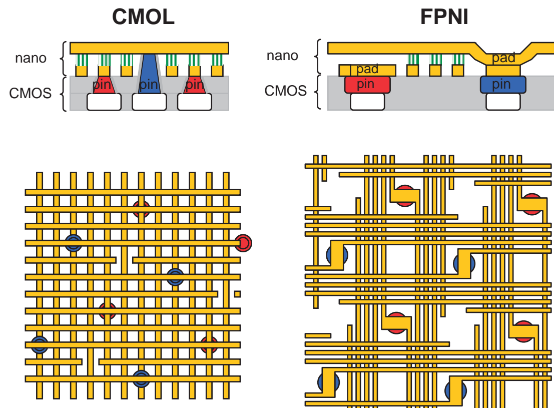


Fig. 4.7 Comparison of CMOL and HP's FPNI circuits (adapted from Ref. [42])

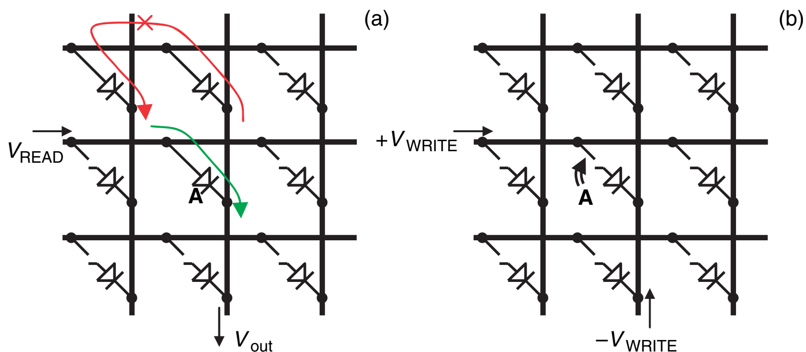


Fig. 4.9 Equivalent circuits of the crossbar memory array showing (**a**) read and (**b**) write operations for one of the cells (marked A). On panel (**a**), *green arrow* shows the useful readout current, while *red arrow* shows the parasitic current to the wrong output wire, which is prevented by the nonlinearity of the $I - V$ curve of device A (if the output voltage is not too high, $V_{out} < V_t$)

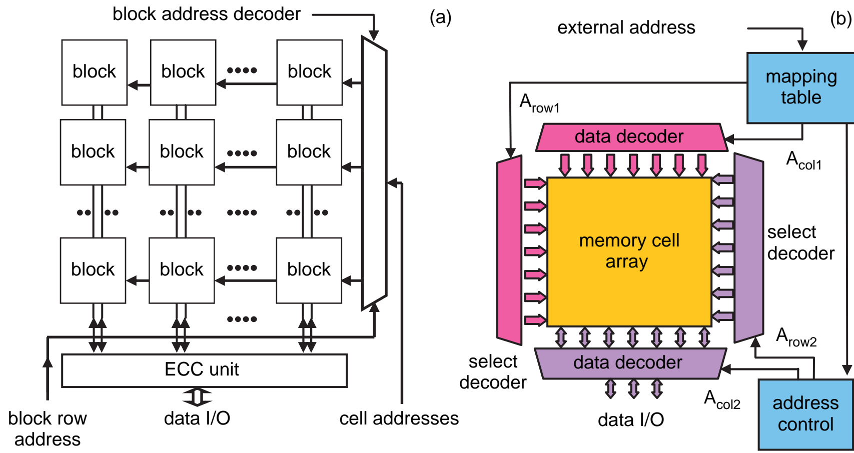


Fig. 4.10 CMOL memory structure: (**a**) global and (**b**) block architectures

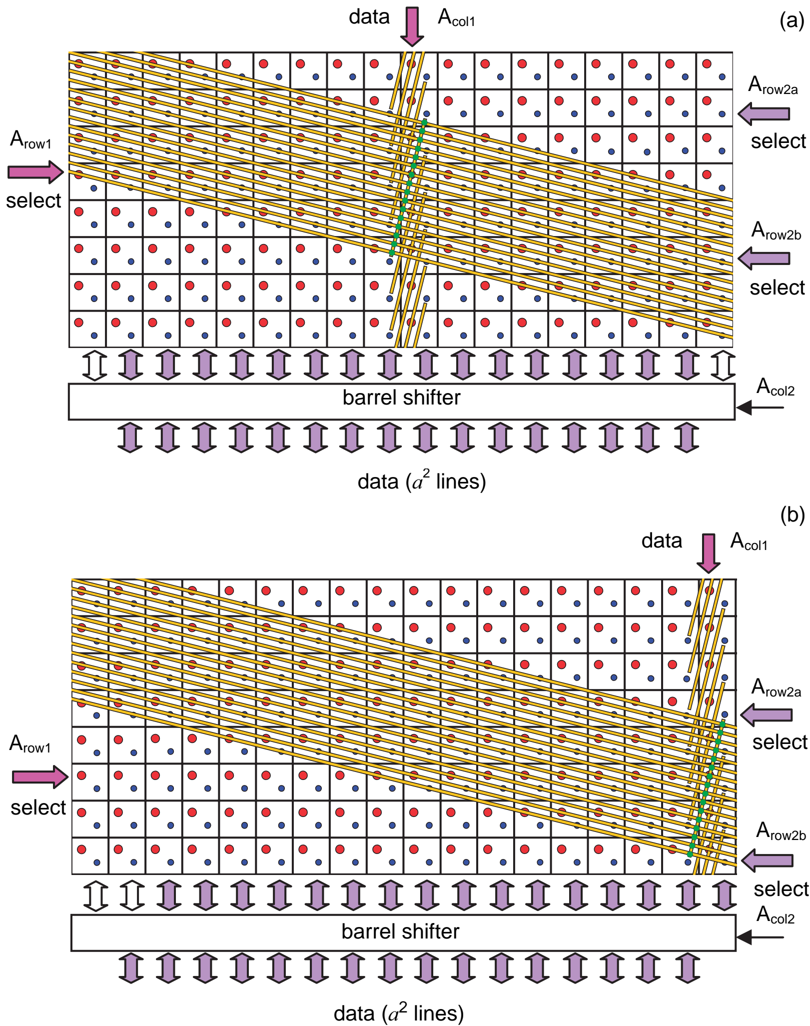


Fig. 4.11 CMOL block architecture: Addressing of an interior column of nanowire segments (for $a = 4$). The figure shows only one (selected) column of the segments, the crosspoint nanodevices connected to one (selected) segment, and continuous top-level nanowires connected to these nanodevices. (In reality, the nanowires of both layers fill all the array plane, with nanodevices at each crosspoint.) The *block arrows* indicate the location of CMOS lines activated at addressing the shown nanodevices

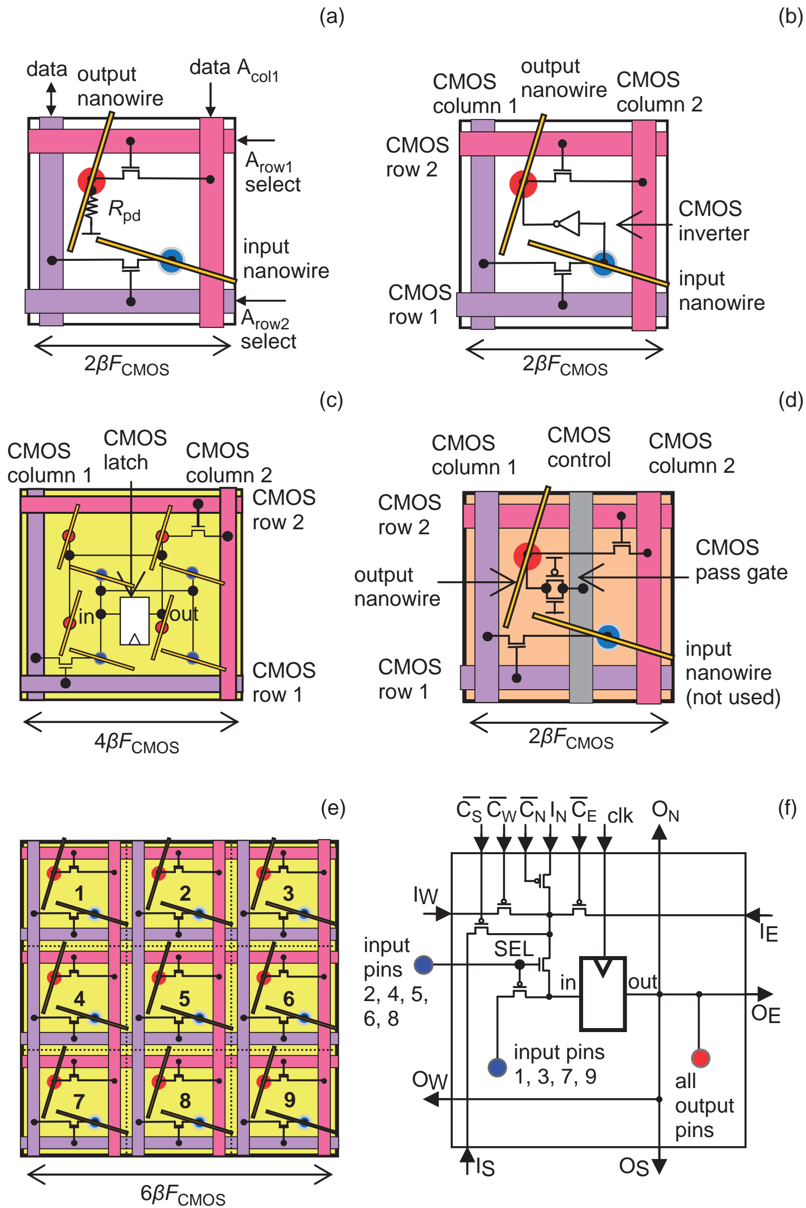


Fig. 4.12 Possible structure of CMOL cells: (**a**) memory relay cell; (**b**) the basic cell: (**c**) the latch cell of CMOL FPGA; (**d**) control cell; (e, f) programmable latch cell of CMOL DSP. Here *red* and *blue points* indicate the corresponding interface pins. For the sake of clarity panels (**a–e**) shows only nanowires which are contacted by interface pins of the given cells. Also for clarity, panel (**e**) shows only the configuration circuitry, while panel (**f**) shows the programmable latch implementation

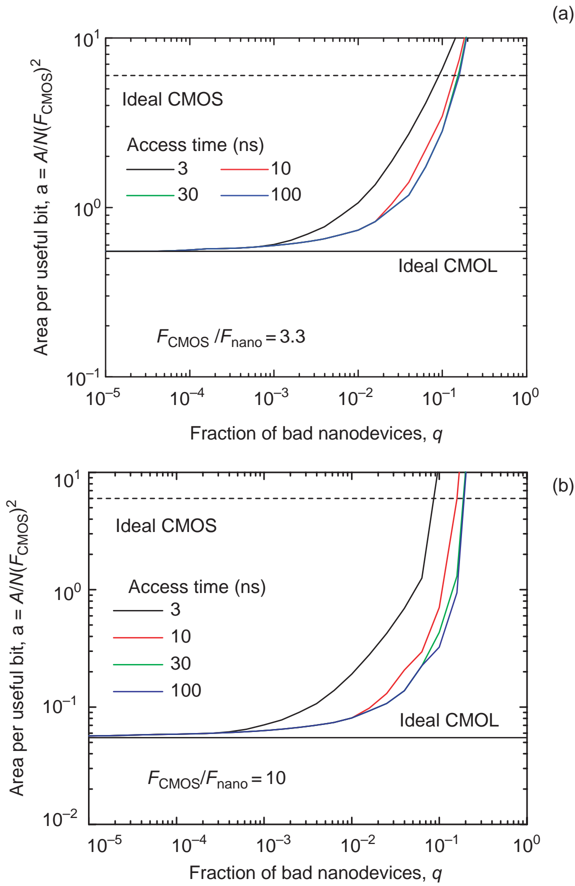


Fig. 4.13 The total chip area per one useful memory cell, as a function of the bad bit fraction q, for several values of the memory access time and two typical values of the F_{CMOS}/F_{nano} ratio. The *horizontal lines* indicate the area for "perfect" CMOS and CMOL memories. In the latter case, this line shows our results for negligible q, while for the former case we use the ITRS data [3] for the densest semiconductor (flash) memories

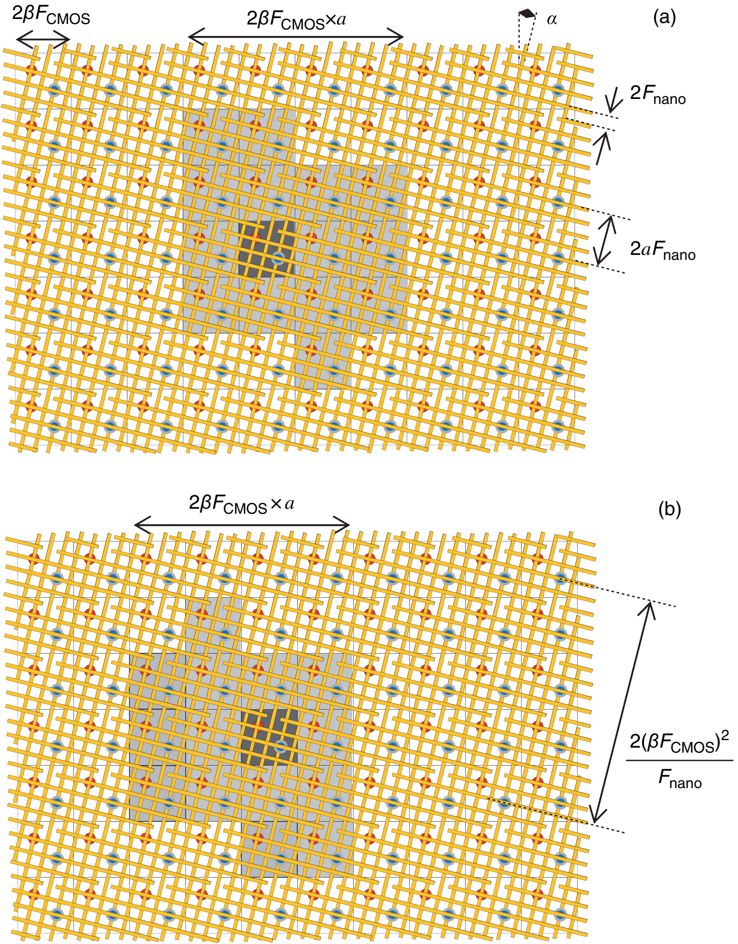


Fig. 4.14 The fragment of one-cell CMOL FPGA fabric for the particular case $a = 4$. In panel (**a**), output pins of $M = a^2 - 2 = 14$ cells (which form the so-called input cell connectivity domain) painted *light gray* may be connected to the input pin of a specific cell (shown *dark gray*) via a pin-nanowire-nanodevice-nanowire-pin links. Similarly, panel (**b**) shows cells (painted *light gray*) whose inputs may be connected directly to the output pin of a specific cell (called output connectivity domain)

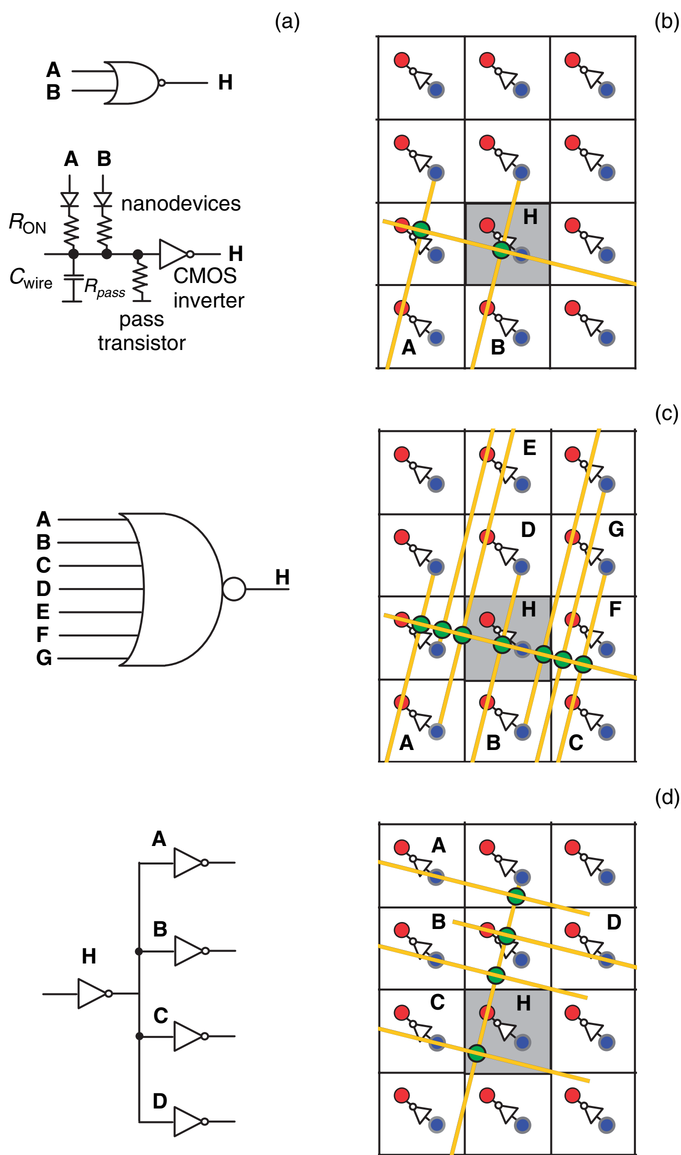


Fig. 4.15 Logic and routing primitives in CMOL FPGA circuits: (**a**) equivalent circuit of fan-in-two NOR gate, (**b**) its physical implementation in CMOL, (**c**) the example of 7-input NOR gate, and (**d**) the example of fan-out of signal to four cells. Note that only several (shown) nanodevices on the input nanowires in panels (**b**), (**c**), and output nanowire in panel (**d**) of cell H are set to the ON state, while others (not shown) are set to the OFF state. Also, for the sake of clarity, panels (**b**)–(**d**) show only the nanowires used for the gate and the broadcast

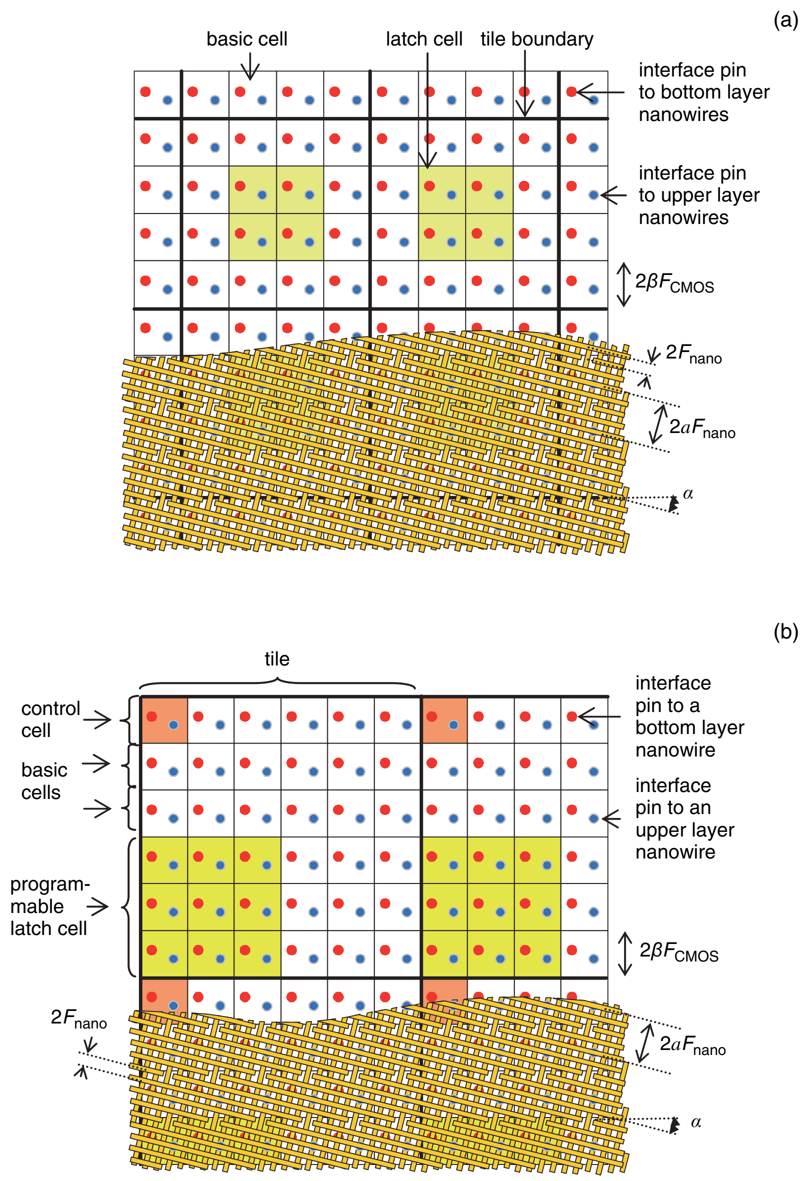


Fig. 4.16 A fragment of (**a**) two-cell CMOL FPGA fabric and (**b**) three-cell CMOL DSP fabric for the particular case $a = 4$

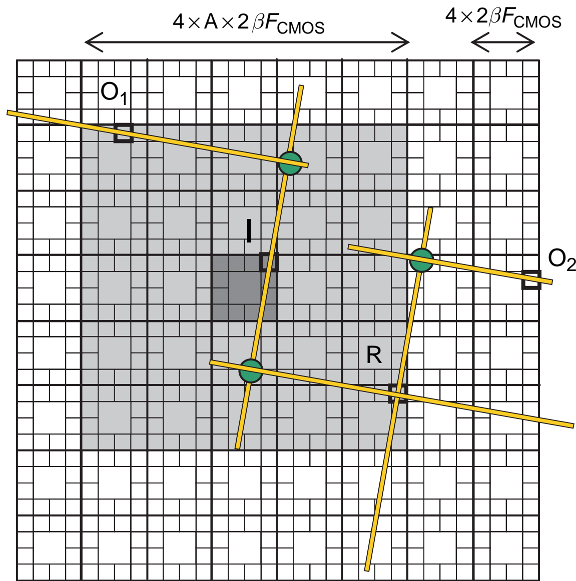


Fig. 4.17 Tile connectivity domain: Any cell of the central tile (shown *dark gray*) can be connected with any cell in the tile connectivity domain (shown *light gray*) via one pin-nanowire-nanodevice-nanowire-pin link (e.g., cells I and O_1). Cells outside of each other's tile connectivity domain (e.g., I and O_2) can be connected with additional routing inverters (e.g., R). Note that nanowire width and nanodevice size are boosted for clarity. For example, for the considered CMOL parameters, 1600 crosspoint nanodevices may fit in one basic cell area

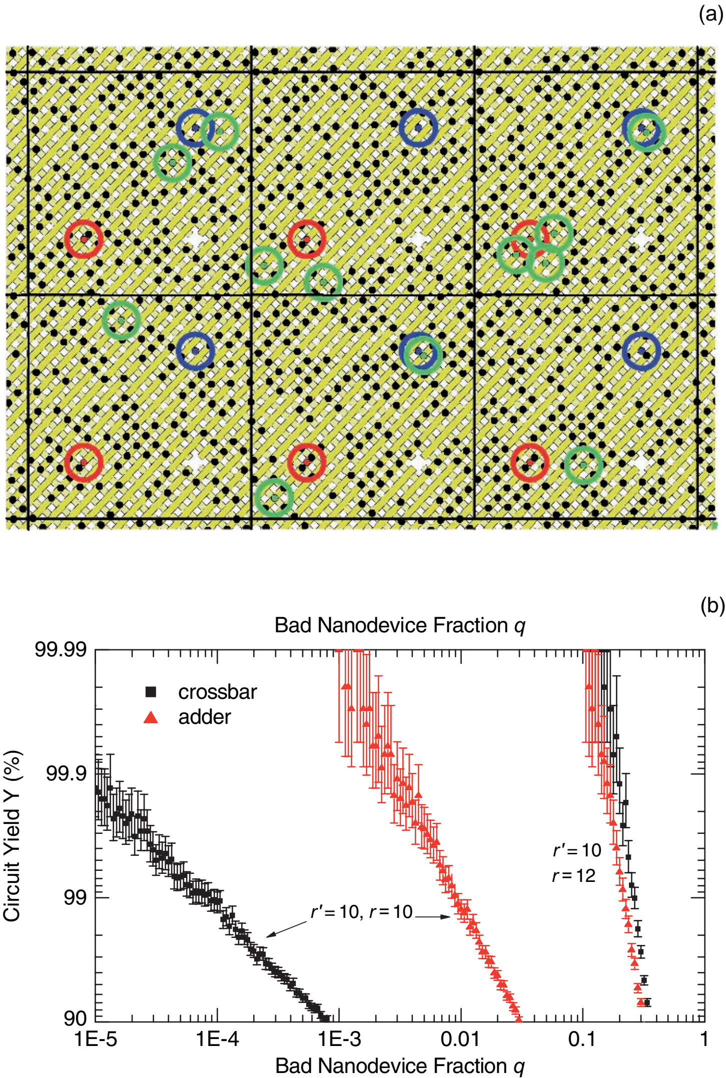


Fig. 4.19 (**a**) A small fragment of the 32-bit Kogge-Stone adder mapped on one-cell CMOL fabric after the reconfiguration as around 50% stuck-on-open nanodevices. Bad nanodevices are shown *black*, good used devices *green*, unused devices are not shown, for clarity. *Colored circles* are only a help for the eye, showing the location of interface pins (*red* and *blue* points) and used nanodevices. *Thin vertical* and *horizontal lines* show CMOS cell borders. (**b**) The final (post-reconfiguration) defect tolerance of 32-bit Kogge-Stone adder and the 64-bit full crossbar for several values of $F_{\mathrm{CMOS}}/F_{\mathrm{nano}}$. For more details – see Ref. [28]

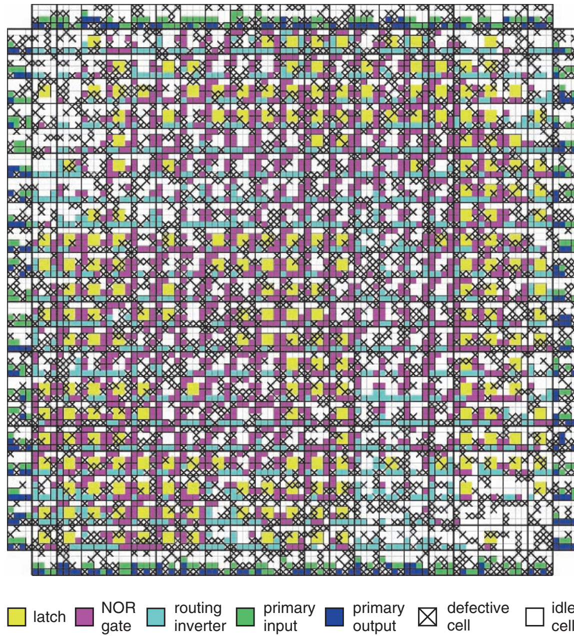


Fig. 4.20 Example of mapping on two-cell CMOL fabric with a presence of defective cells: dsip.blif circuit of the Toronto 20 set, mapped on the $(21+2)\times(21+2)$ tile array with 30% defective cells. Here the additional layer of tiles at the array periphery is used exclusively for I/O functions. The cells from these peripheral tiles are functionally similar to input and output pads and cannot be configured to NOR gates

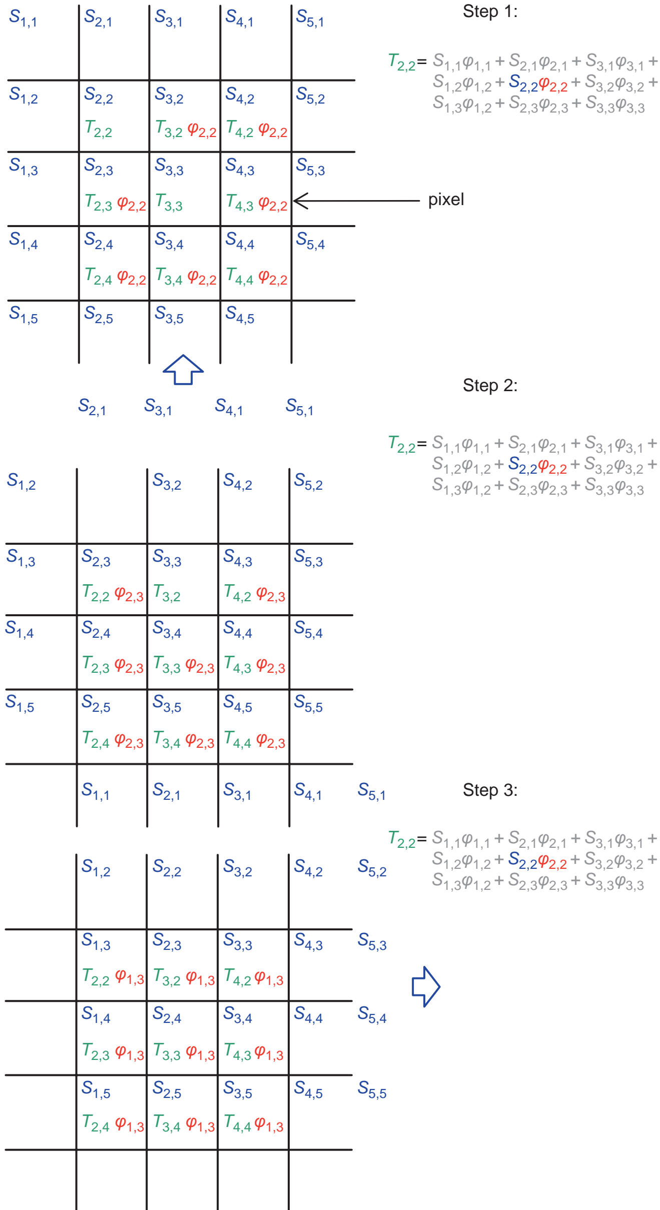


Fig. 4.23 Three sequential time steps of the convolution in the *left top* corner of the CMOL DSP array for $F = 3$. Colored terms in the formulas below each panel show the calculated partial sums in the pixel 2,2. For the (uncharacteristically small) filter size, it takes just $F^2 = 9$ steps to complete the processing of one frame

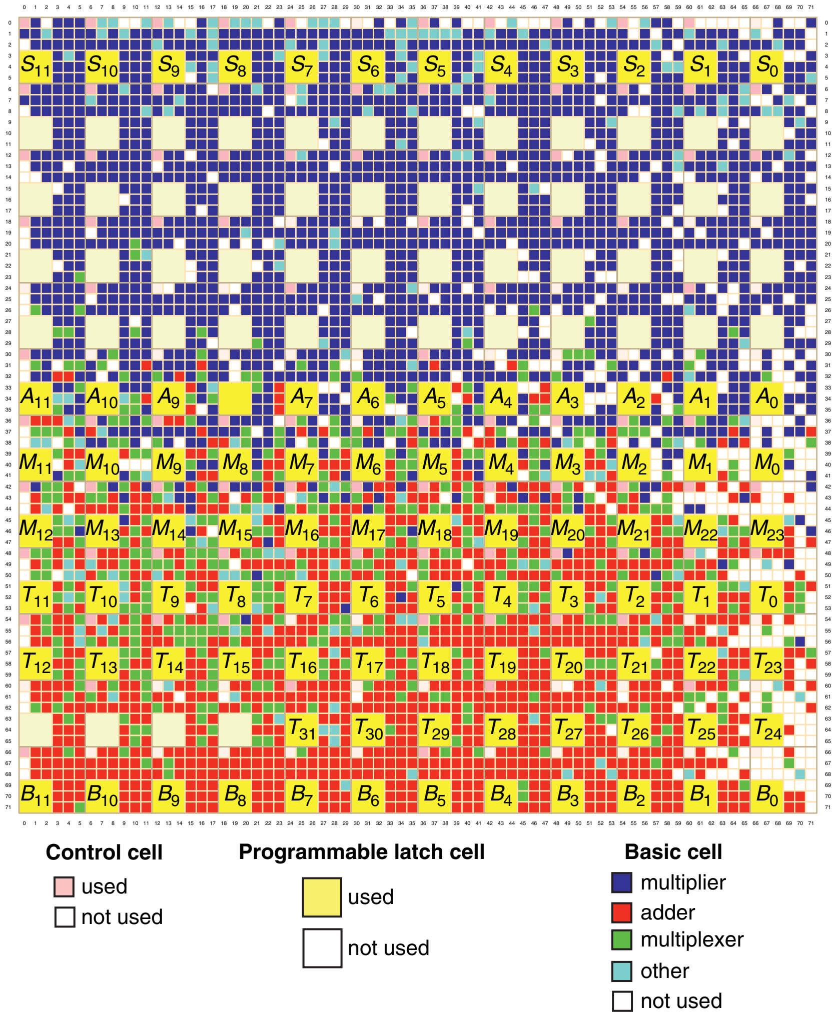


Fig. 4.25 The mapping of the pixel on CMOL DSP (for $F_{\mathrm{CMOS}}/F_{\mathrm{nano}} = 10$) after its successful reconfiguration of the circuit around as many as 40% of bad nanodevices with random locations. Programmable latches A and B are used for bypass circuitry during the data up and down shift operations

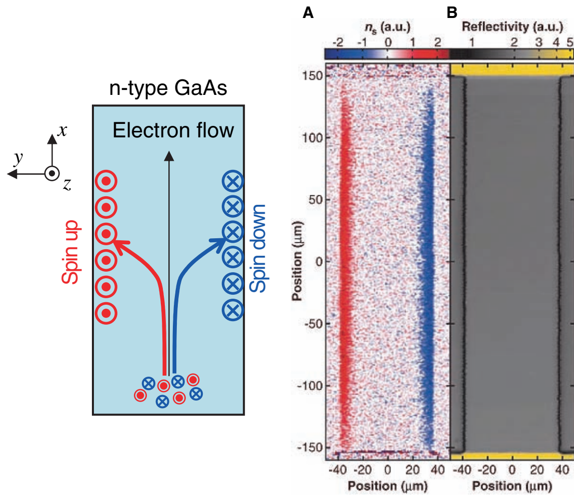
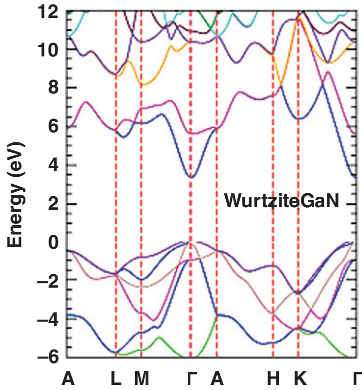
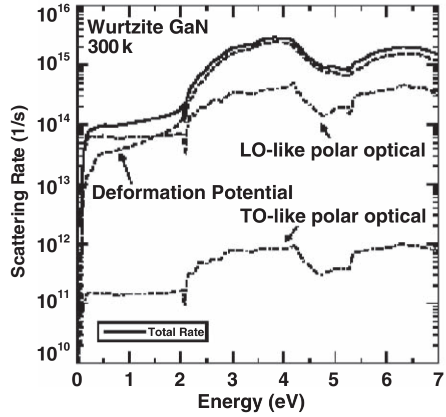


Fig. 5.35 (*Left*) General behavior of the spin Hall effect, where an unpolarized charge current generates a transverse spin current. (*Right*) Magneto-optical imaging of the spin Hall effect in n-type GaAs, from Ref. [159]. Reprinted with permission from AAAS





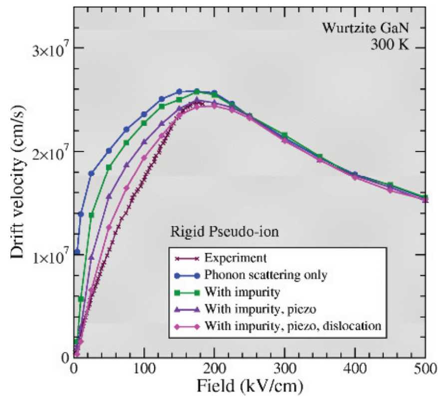


Fig. 6.8 Calculated bandstructure (**a**), rigid-ion scattering rates (**b**), and calculated velocity-field characteristics (**c**) from the CMC simulator for Wurzite GaN at 300 K. Reprinted with permission from Ref. [39], Copyright 2004, Institute of Physics Publishing

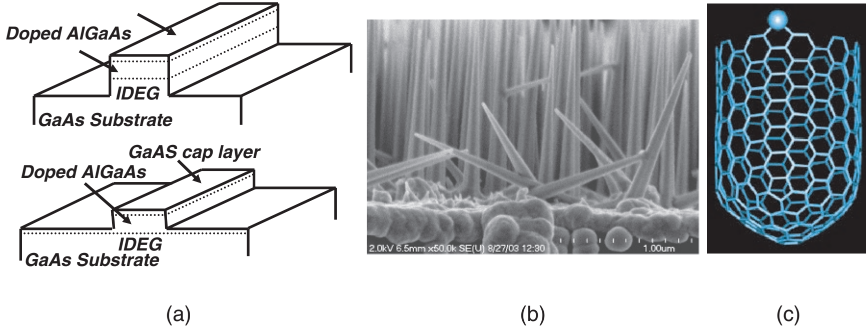


Fig. 6.14 Different experimental realizations of quasi-1D systems. (**a**) Different structures realized through lateral etching or confinement of a 2D quantum well structure. (**b**) A self-assembled Si nanowire structure grown using vapour–liquid–solid epitaxy. (**c**) A carbon nanotube

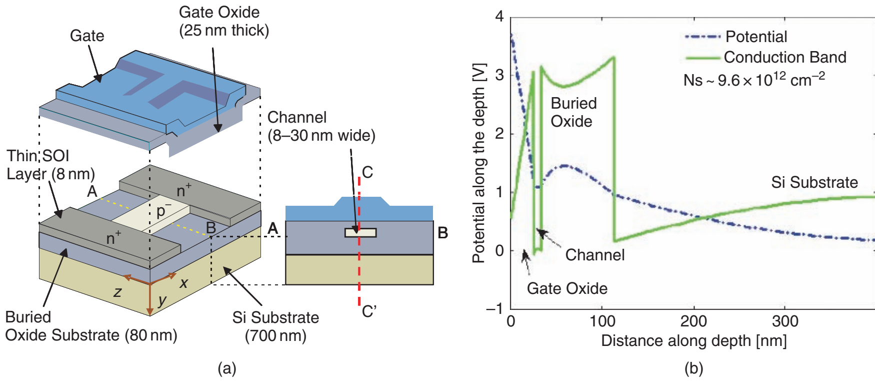


Fig. 6.15 The *left panel* (**a**) shows the schematic of a simulated SiNW on ultrathin SOI. The conduction band profile on the *right side* (**b**) is taken along the red cutline CC from the top panel. The width of the channel is 30 nm [53]

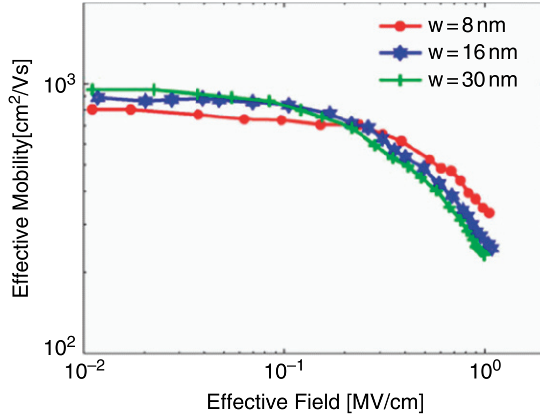


Fig. 6.16 Variation of the field-dependent mobility with varying SiNW width. The wire thickness is kept constant at 8 nm [53]

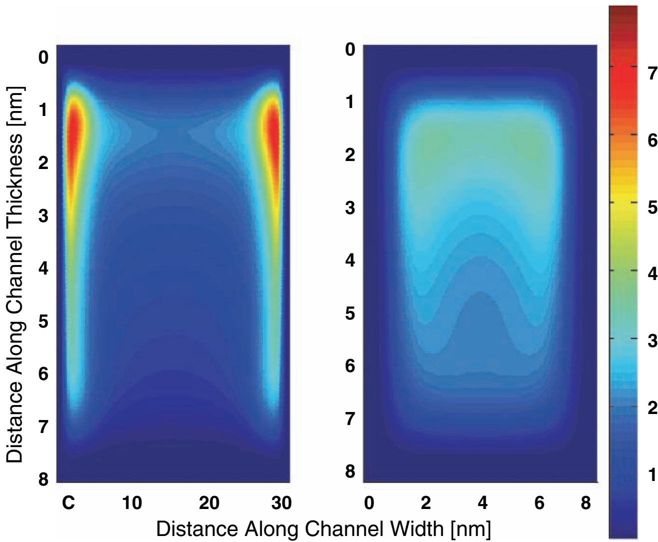


Fig. 6.17 Electron distribution across the nanowire, for the wire width of 30 nm (*left panel*) and 8 nm (*right panel*). In both panels, the transverse field is 1 MV/cm, the wire thickness is 8 nm, and the color scale is in $\times 10^{19}\ \text{cm}^{-3}$ [53]

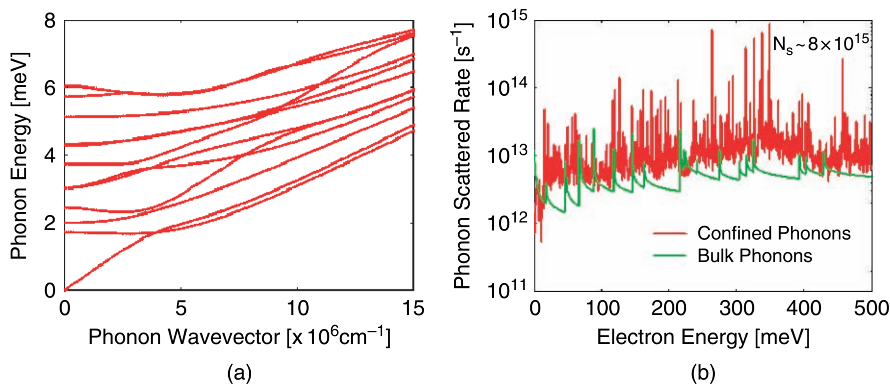


Fig. 6.18 Effect of confinement on the phonon dispersion in a SiNW (**a**), and the corresponding effect on the quasi-1D scattering rate (**b**)

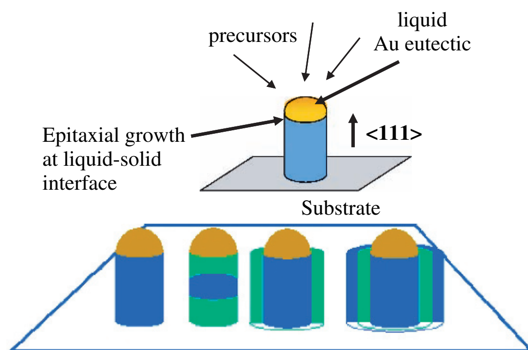


Fig. 6.19 Schematic of growth of a semiconductor nanowires using vapor–liquid– solid (VLS) phase growth. The *bottom panel* illustrates several different heterostructures realizable using this technique

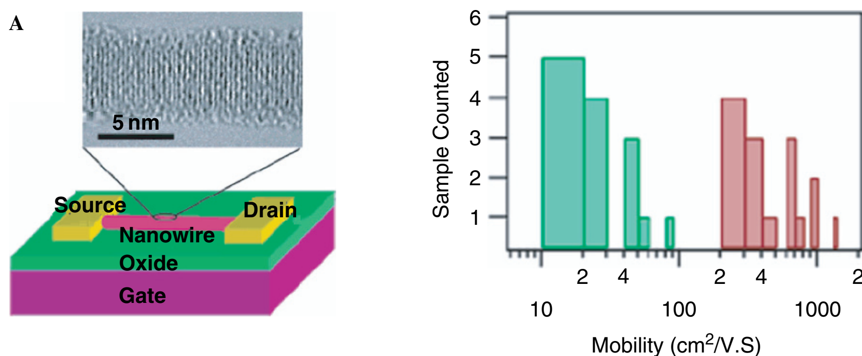


Fig. 6.20 Si Nanowire field effect transistor structure. The *left panel* shows a schematic and electron micrograph of the transistor structure. The *right panel* shows the measured mobility before (green data, left side) and after (pink data, right side) surface modification. Reprinted with permission from Ref [47], Copyright 2003

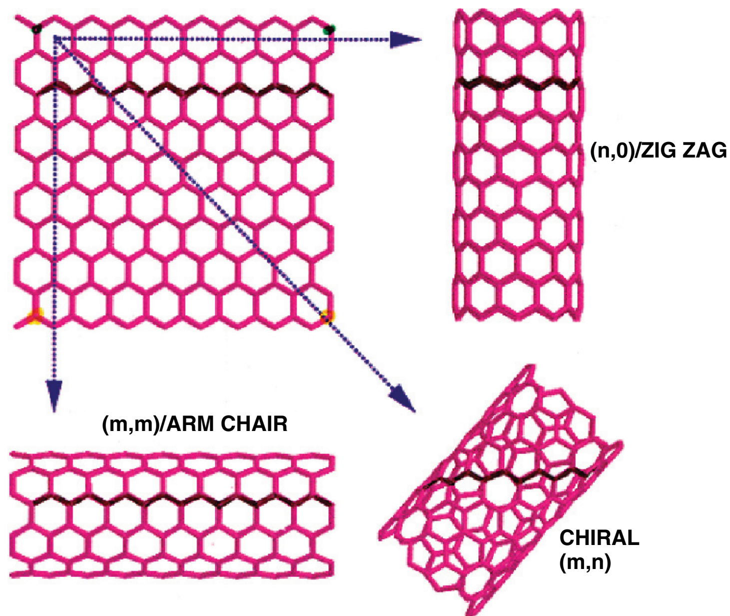


Fig. 6.21 Molecular structure of a single-wall CNT, formed by rolling a sheet of graphene, illustrating different chiralities

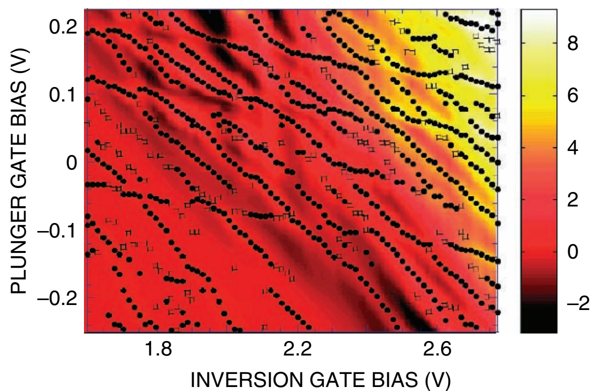


Fig. 6.37 Conductance peak positions as function of both inversion gate and plunger gate bias exhibiting crossing and anti-crossing behaviors of apparent level structure of dot [103]

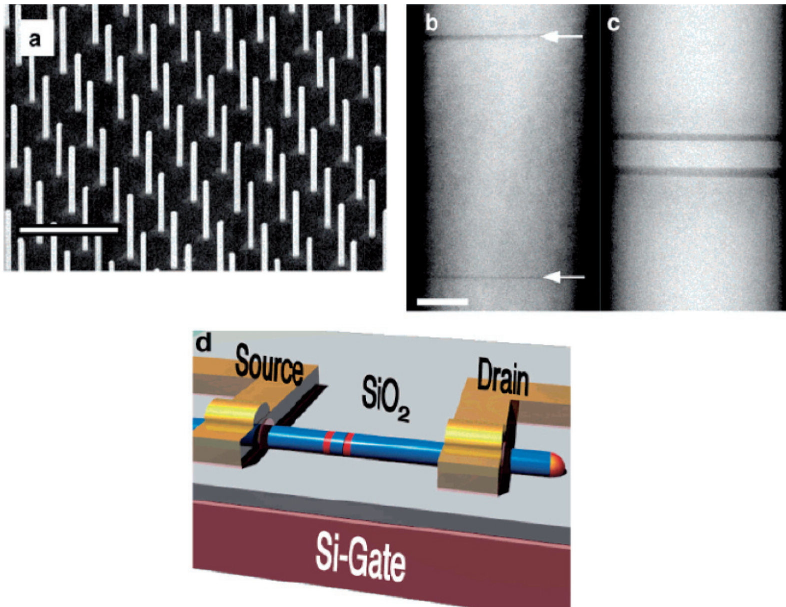


Fig. 6.38 Characterization and processing of nanowires. (**a**) Scanning electron micrograph of homogeneous InAs nanowires grown on an InAs substrate from lithographically defined arrays of Au particles. The image demonstrates the ability of the CBE to produce identical nanowire devices. The scale bar corresponds to 1 μm. (**b**) Dark-field scanning transmission electron microscopy image of a nanowire with a 100 nm long InAs quantum dot between two very thin InP barriers. Scale bar depicts 20 nm. (**c**) Corresponding image of a 10 nm long InAs dot. The InP barrier thickness is 3 and 3.7 nm, respectively. (**d**) The heterostructured wires are deposited on a SiO_2-capped Si substrate and source and drain contacts are fabricated by lithography. Reprinted with permission from Ref. [102], Copyright 2004, American Chemical Society

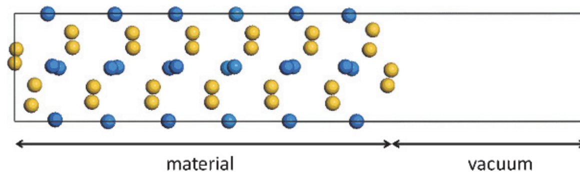


Fig. 7.2 Supercell used to simulate the (101) surface of PtSi. Si (Pt) atoms are represented with *yellow* (*blue*) color. The [101] direction is along the long side of the supercell

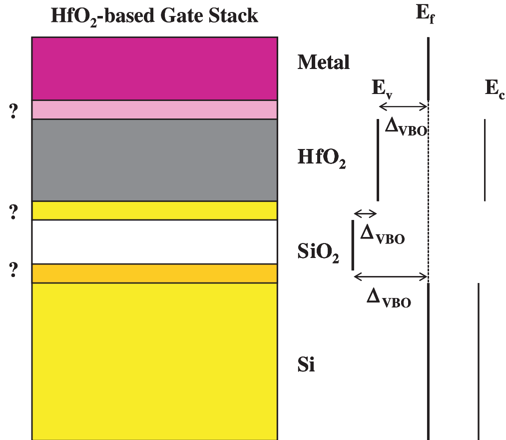


Fig. 7.3 Schematic of a multilayer gate stack based on HfO_2. A plausible band alignment across the stack is also shown

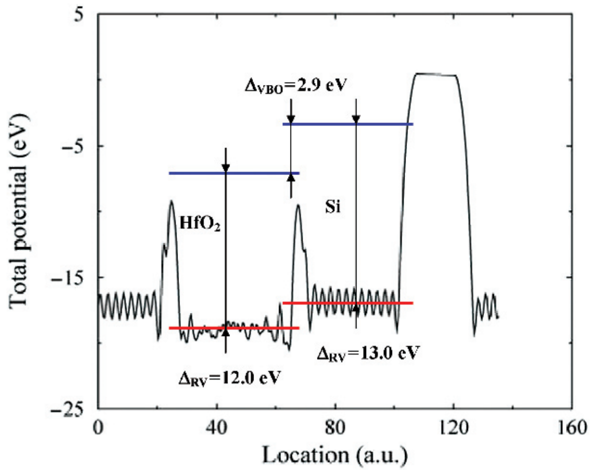


Fig. 7.4 The valence band offset between Si and HfO_2 is calculated using the reference potential method. The average reference potential is indicated with *red lines*, and the valence band maxima with *blue lines*. The discontinuity is estimated to be 2.9 eV

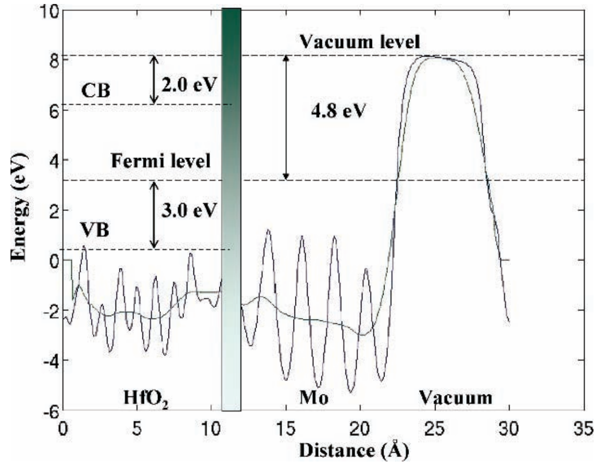


Fig. 7.6 The average reference potential across the Mo–HfO_2 heterojunction. The vacuum level, conduction and valence band of HfO_2 and the Fermi level are indicated

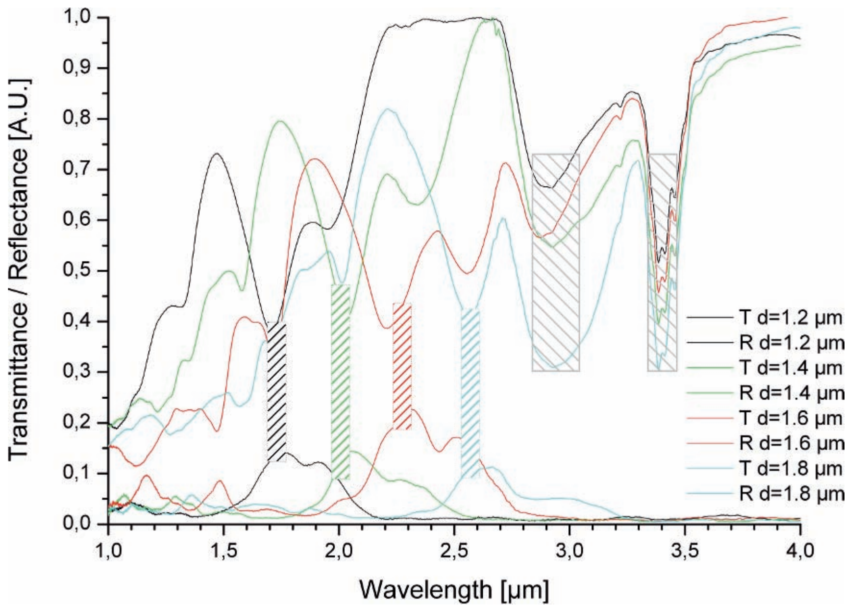


Fig. 12.11 Transmission and reflection spectra of woodpile photonic crystals with different rod spacings