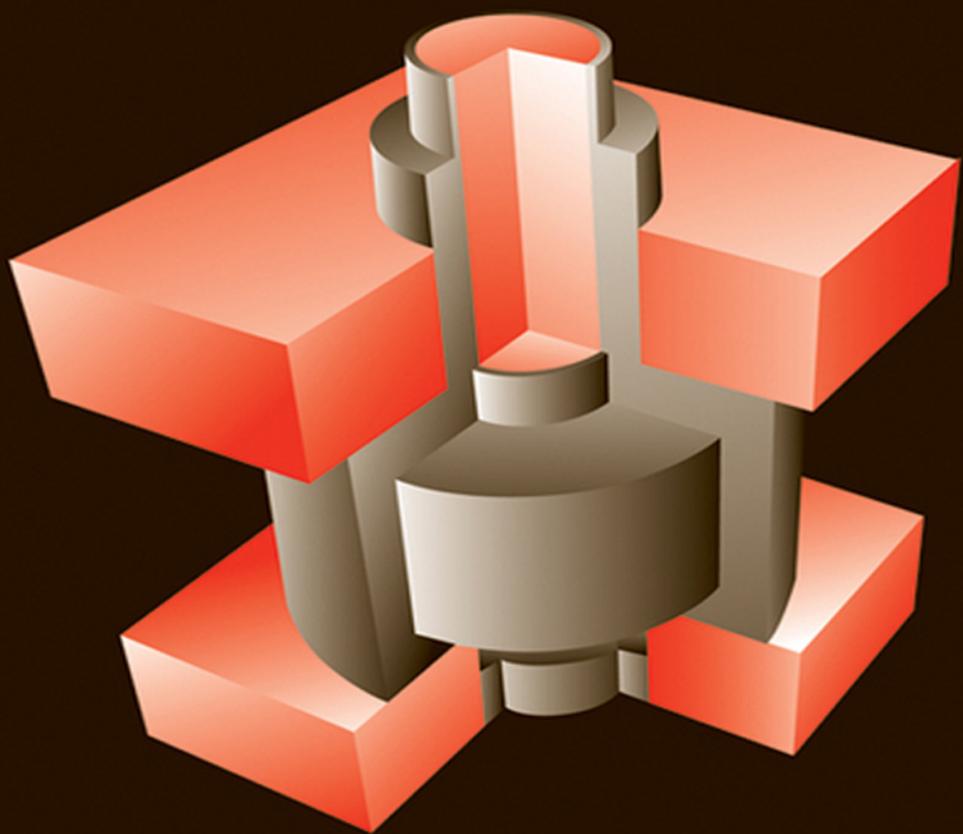


Semiconductor Devices

Physics and Technology

S. M. SZE

M. K. LEE



3RD
EDITION

3RD EDITION

Semiconductor Devices

Physics and Technology

S. M. SZE

*EtronTech Distinguished Chair Professor
College of Electrical and Computer Engineering
National Chiao Tung University
Hsinchu, Taiwan*

M. K. LEE

*Professor
Department of Electrical Engineering
National Sun Yat-sen University
Kaohsiung, Taiwan*



JOHN WILEY & SONS, INC.

Acquisitions Editor *Dan Sayre*
Marketing Manager *Christopher Ruel*
Senior Editorial Assistant *Katie Singleton*
Editorial Program Assistant *Samantha Mendel*
Production Manager *Micheline Frederick*
Cover Designer *Wendy Lai*
Pre-press Service *Robots & Cupcakes*

This book was typeset in *Times Roman* by the authors and printed and bound by *Quad Graphics/Versailles*. The cover was printed by *Quad Graphics/Versailles*.

cover photo: © 2010 IEEE. Reprinted, with permission, from IEDM Technical Digest, S. Whang et. al, "Novel 3-dimensional Dual Control-gate with Surrounding Floating-gate (DC-SF) NAND flash cell for 1Tb file storage application."

The book is printed on acid-free paper. 

Copyright © 1985, 2002, 2012 by John Wiley & Sons, Inc. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (508) 750-8400, fax (508) 750-4470. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc. 605 Third Avenue, New York, NY 10158-0012, (212) 850-6008, E-mail: PERMREQ@WILEY.COM. To order books or for customer service call 1-800-CALL-WILEY (225-5945).

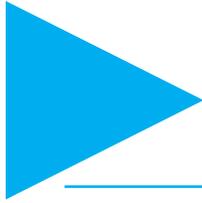
Evaluation copies are provided to qualified academics and professionals for review purposes only, for use in their courses during the next academic year. These copies are licensed and may not be sold or transferred to a third party. Upon completion of the review period, please return the evaluation copy to Wiley. Return instructions and a free of charge return shipping label are available at www.wiley.com/go/returnlabel. Outside of the United States, please contact your local representative.

ISBN 978-0470-53794-7

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

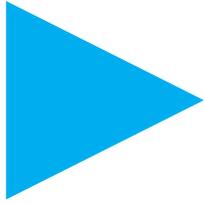
In memory of
Prof. John L. Moll (1921 ~ 2011)
A pioneer of Semiconductor Devices.
Also a big thanks to thepiratebay.se
for hosting the torrent of this book for
the whole world to use for free!
Praise Jesus



Contents

| | | |
|--|--|------------|
| Preface | | |
| Acknowledgments | | |
| ▶ CHAPTER 0 | | |
| Introduction | | |
| 0.1 Semiconductor Devices | | |
| 0.2 Semiconductor Technology | | |
| Summary | | |
| <hr/> | | |
| PART I | | |
| SEMICONDUCTOR PHYSICS | | |
| <hr/> | | |
| ▶ CHAPTER 1 | | |
| Energy Bands and Carrier Concentration in Thermal Equilibrium | | |
| 1.1 Semiconductor Materials | | |
| 1.2 Basic Crystal Structures | | |
| 1.3 Valence Bonds | | |
| 1.4 Energy Bands | | |
| 1.5 Intrinsic Carrier Concentration | | |
| 1.6 Donors and Acceptors | | |
| Summary | | |
| ▶ CHAPTER 2 | | |
| Carrier Transport Phenomena | | |
| 2.1 Carrier Drift | | |
| 2.2 Carrier Diffusion | | |
| 2.3 Generation and Recombination Processes | | |
| 2.4 Continuity Equation | | |
| 2.5 Thermionic Emission Process | | |
| 2.6 Tunneling Process | | |
| 2.7 Space-Charge Effect | | |
| 2.8 High-Field Effects | | |
| Summary | | |
| <hr/> | | |
| PART II | | |
| SEMICONDUCTOR DEVICES | | |
| <hr/> | | |
| ▶ CHAPTER 3 | | |
| <i>p-n</i> Junction | | |
| 3.1 Thermal Equilibrium Condition | | |
| 3.2 Depletion Region | | |
| 3.3 Depletion Capacitance | | |
| 3.4 Current-Voltage Characteristics | | |
| 3.5 Charge Storage and Transient Behavior | | |
| 3.6 Junction Breakdown | | |
| 3.7 Heterojunction | | |
| Summary | | |
| vii | ▶ CHAPTER 4 | |
| ix | Bipolar Transistors and Related Devices | 123 |
| 1 | 4.1 Transistor Action | 124 |
| | 4.2 Static Characteristics of Bipolar Transistors | 129 |
| 1 | 4.3 Frequency Response and Switching of Bipolar Transistors | 137 |
| 6 | 4.4 Nonideal Effects | 142 |
| 12 | 4.5 Heterojunction Bipolar Transistors | 146 |
| | 4.6 Thyristors and Related Power Devices | 149 |
| | Summary | 155 |
| | ▶ CHAPTER 5 | |
| | MOS Capacitor and MOSFET | 160 |
| 15 | 5.1 Ideal MOS Capacitor | 160 |
| 15 | 5.2 SiO ₂ -Si MOS Capacitor | 169 |
| 17 | 5.3 Carrier Transport in MOS Capacitors | 174 |
| 22 | 5.4 Charge-Coupled Devices | 177 |
| 23 | 5.5 MOSFET Fundamentals | 180 |
| 29 | Summary | 192 |
| 34 | ▶ CHAPTER 6 | |
| 40 | Advanced MOSFET and Related Devices | 195 |
| | 6.1 MOSFET Scaling | 195 |
| 43 | 6.2 CMOS and BiCMOS | 205 |
| 43 | 6.3 MOSFET on Insulator | 210 |
| 53 | 6.4 MOS Memory Structures | 214 |
| 56 | 6.5 Power MOSFET | 223 |
| 62 | Summary | 224 |
| 68 | ▶ CHAPTER 7 | |
| 69 | MESFET and Related Devices | 228 |
| 71 | 7.1 Metal-Semiconductor Contacts | 229 |
| 73 | 7.2 MESFET | 240 |
| 77 | 7.3 MODFET | 249 |
| | Summary | 255 |
| | ▶ CHAPTER 8 | |
| | Microwave Diodes; Quantum-Effect and Hot-Electron Devices | 258 |
| 82 | 8.1 Microwave Frequency Bands | 259 |
| 83 | 8.2 Tunnel Diode | 260 |
| 87 | 8.3 IMPATT Diode | 260 |
| 95 | 8.4 Transferred-Electron Devices | 265 |
| 99 | 8.5 Quantum-Effect Devices | 269 |
| 108 | 8.6 Hot-Electron Devices | 274 |
| 111 | Summary | 277 |
| 117 | | |
| 120 | | |

| | | | |
|--|------------|-------------------------------------|-----|
| ▶ CHAPTER 9 | | | |
| Light Emitting Diodes and Lasers | 280 | | |
| 9.1 Radiative Transitions and Optical Absorption | 280 | | |
| 9.2 Light-Emitting Diodes | 286 | | |
| 9.3 Various Light-Emitting Diodes | 291 | | |
| 9.4 Semiconductor Lasers | 302 | | |
| Summary | 319 | | |
| ▶ CHAPTER 10 | | | |
| Photodetectors and Solar Cells | 323 | | |
| 10.1 Photodetectors | 323 | | |
| 10.2 Solar Cells | 336 | | |
| 10.3 Silicon and Compound-Semiconductor Solar Cells | 343 | | |
| 10.4 Third-Generation Solar Cells | 348 | | |
| 10.5 Optical Concentration | 352 | | |
| Summary | 352 | | |
| PART III | | | |
| SEMICONDUCTOR TECHNOLOGY | | | |
| ▶ CHAPTER 11 | | | |
| Crystal Growth and Epitaxy | 357 | | |
| 11.1 Silicon Crystal Growth from the Melt | 357 | | |
| 11.2 Silicon Float-Zone Process | 363 | | |
| 11.3 GaAs Crystal-Growth Techniques | 367 | | |
| 11.4 Material Characterization | 370 | | |
| 11.5 Epitaxial-Growth Techniques | 377 | | |
| 11.6 Structures and Defects in Epitaxial Layers | 384 | | |
| Summary | 388 | | |
| ▶ CHAPTER 12 | | | |
| Film Formation | 392 | | |
| 12.1 Thermal Oxidation | 392 | | |
| 12.2 Chemical Vapor Deposition of Dielectrics | 400 | | |
| 12.3 Chemical Vapor Deposition of Polysilicon | 409 | | |
| 12.4 Atom Layer Deposition | 412 | | |
| 12.5 Metallization | 414 | | |
| Summary | 425 | | |
| ▶ CHAPTER 13 | | | |
| Lithography and Etching | 428 | | |
| 13.1 Optical Lithography | 428 | | |
| 13.2 Next-Generation Lithographic Methods | 441 | | |
| 13.3 Wet Chemical Etching | 447 | | |
| 13.4 Dry Etching | 450 | | |
| Summary | 462 | | |
| ▶ CHAPTER 14 | | | |
| Impurity Doping | 466 | | |
| 14.1 Basic Diffusion Process | 467 | | |
| 14.2 Extrinsic Diffusion | 476 | | |
| 14.3 Diffusion-Related Processes | 480 | | |
| | | 14.4 Range of Implanted Ions | 483 |
| | | 14.5 Implant Damage and Annealing | 490 |
| | | 14.6 Implantation-Related Processes | 495 |
| | | Summary | 501 |
| ▶ CHAPTER 15 | | | |
| Integrated Devices | 505 | | |
| 15.1 Passive Components | 507 | | |
| 15.2 Bipolar Technology | 511 | | |
| 15.3 MOSFET Technology | 516 | | |
| 15.4 MESFET Technology | 529 | | |
| 15.5 Challenges for Nanoelectronics | 532 | | |
| Summary | 537 | | |
| ▶ APPENDIX A | | | |
| List of Symbols | | | 541 |
| ▶ APPENDIX B | | | |
| International Systems of Units (SI Units) | | | 543 |
| ▶ APPENDIX C | | | |
| Unit Prefixes | | | 544 |
| ▶ APPENDIX D | | | |
| Greek Alphabet | | | 545 |
| ▶ APPENDIX E | | | |
| Physical Constants | | | 546 |
| ▶ APPENDIX F | | | |
| Properties of Important Element and Binary Compound Semiconductors at 300 K | | | 547 |
| ▶ APPENDIX G | | | |
| Properties of Si and GaAs at 300 K | | | 548 |
| ▶ APPENDIX H | | | |
| Derivation of the Density of States in a Semiconductor | | | 549 |
| ▶ APPENDIX I | | | |
| Derivation of Recombination Rate for Indirect Recombination | | | 553 |
| ▶ APPENDIX J | | | |
| Calculation of the Transmission Coefficient for a Symmetric Resonant-Tunneling Diode | | | 555 |
| ▶ APPENDIX K | | | |
| Basic Kinetic Theory of Gases | | | 557 |
| ▶ APPENDIX L | | | |
| Answers to Selected Problems | | | 559 |
| Photo credits | | | 563 |
| Index | | | 565 |



Preface

The book is an introduction to the physical principles of modern semiconductor devices and their advanced fabrication technology. It is intended as a text for undergraduate students in applied physics, electrical and electronics engineering, and materials science. It can also serve as a reference for graduate students and practicing engineers as well as scientists who are not familiar with the subject or need an update on device and technology developments.

▶ WHAT'S NEW IN THE THIRD EDITION

- 35% of the material has been revised or updated. We have added many sections of current interest such as CMOS image sensors, FinFET, 3rd generation solar cells, and atomic layer deposition. In addition, we have omitted or reduced sections of less important topics to maintain the overall book length.
- We have expanded the treatment of MOSFET and related devices to two chapters because of their importance in electronic applications. We have also expanded the treatment of photonic devices to two chapters because of their importance in communication and alternative energy sources.
- To improve the development of each subject, sections that contain graduate-level mathematics or physical concepts have been omitted or moved to the Appendixes.

▶ TOPICAL COVERAGE

- Chapter 0 gives a brief historical review of major semiconductor devices and key technology developments. The following text is organized in three parts.
- Part I, Chapters 1–2, describes the basic properties of semiconductors and their conduction processes, with special emphasis on the two most important semiconductors, silicon (Si) and gallium arsenide (GaAs). The concepts in Part I, which will be used throughout the book, require a background knowledge of modern physics and college calculus.
- Part II, Chapters 3–10, discusses the physics and characteristics of all major semiconductor devices. We begin with the p–n junction, the key building block of most semiconductor devices. We proceed to bipolar and field-effect devices and then cover microwave, quantum-effect, hot-electron, and photonic devices.
- Part III, Chapters 11–15, deals with processing technology from crystal growth to impurity doping. We present the theoretical and practical aspects of the major steps in device fabrication with an emphasis on integrated devices.

► KEY FEATURES

Each chapter includes the following features:

- The chapter starts with an overview of the topical contents. A list of covered learning goals is also provided.
- The third edition contains many worked-out examples that apply basic concepts to specific problems.
- A chapter summary at the end of each chapter summarizes the important concepts and helps the student review the content before tackling the homework problems that follow.
- The book includes about 250 homework problems. Answers to odd-numbered problems with numerical solutions are provided in Appendix L.

► COURSE DESIGN OPTIONS

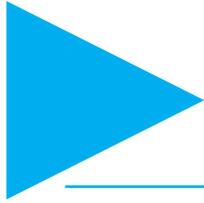
The third edition can provide greater flexibility in course design. The book contains enough material for a full-year sequence in device physics and processing technology. Assuming three lectures per week, a two-semester sequence can cover Chapters 0–7 in the first semester, leaving Chapters 8–15 for the second semester. For a three-quarter sequence, the logical breakpoints are Chapters 0–5, Chapters 6–10, and Chapters 11–15.

A two-quarter sequence can cover Chapters 0–5 in the first quarter. The instructor has several options for the second quarter. For example, covering Chapters 6, 12, 13, 14 and 15 produces a strong emphasis on MOSFET and related process technologies, while covering Chapters 6–10 emphasizes all major devices. For a one-quarter course on semiconductor device processing, the instructor can cover Section 0.2 and Chapters 11–15.

A one-semester course on basic semiconductor physics and devices can cover Chapters 0–7. A one-semester course on microwave and photonic devices can cover Chapters 0–3, and 7–10. For students with some familiarity with semiconductor fundamentals, a one-semester course on MOSFET physics and technology can cover Chapters 0, 5, 6, and 11–15. Of course, there are many other course design options depending on the teaching schedule and the instructor's choice of topics.

► TEXTBOOK SUPPLEMENTS

- Instructor's Manual. A complete set of detailed solutions to all the end-of-chapter problems has been prepared. These solutions are available free to all adopting faculty.
- The figures used in the text are available to instructors in electronic format, from the publisher. More information is available at the publisher's website: <http://www.wiley.com/college/sze>



Acknowledgments

In the course of writing the text, we had the good fortune of help and support from many people. First we express our gratitude to the management of our academic institutions, the National Chiao Tung University and the National Sun Yat-sen University, without whose support this book could not have been written. One of us (S. M. Sze) would like to thank Etron Technology Inc., Taiwan, ROC, for the EtronTech Distinguished Chair Professorship grant that provided the environment to work on this book.

Many people have assisted us in revising this book. We have benefited significantly from suggestions made by the reviewers who took time from their busy schedules for careful scrutiny of this book. Credit is due to the following scholars: Prof. C. C. Chang of the National Taiwan Ocean University; Profs. L. B. Chang and C. S. Lai of the Chang Gung University; Dr. O. Cheng and Mr. T. Kao of the United Microelectronics Corporation (UMC); Dr. S. C. Chang and Dr. Y. L. Wang of the Taiwan Semiconductor Manufacturing Company (TSMC); Prof. T. C. Chang of the National Sun Yat-sen University; Profs. T. S. Chao, H. C. Lin, P. T. Liu, and T. Wang of the National Chiao Tung University; Prof. J. Gong of the Tunghai University; Profs. C. F. Huang and M. C. Wu of the National Tsing Hua University; Profs. C. J. Huang and W. K. Yeh of the National University of Kaohsiung; Profs. J. G. Hwu, C. Liu, and L. H. Peng of the National Taiwan University; Prof. J. W. Hong of the National Central University; Profs. W. C. Hsu and W. C. Liu of the National Cheng Kung University; Profs. Y. L. Jiang and D. S. Wu of the National Chung Hsing University; Prof. C. W. Wang of the National Chung Cheng University; Dr. C. L. Wu of Transcom. Inc.; and Dr. Y. H. Yang of PixArt Imaging Inc.

We are further indebted to Mr. N. Erdos for technical editing of the manuscript. In each case where an illustration was used from another published source, we have received permission from the copyright holder. Even though all illustrations were then adapted and redrawn, we appreciate being granted these permissions. At John Wiley & Sons, we wish to thank Mr. D. Sayre and Mr. G. Telecki, who encouraged us to undertake the project. One of us (M. K. Lee) would like to thank his daughter Ko-Hui for preparing homework problems and solutions. Finally, we are grateful to our wives, Therese Sze and Amanda Lee, for their support and assistance over the course of the book project.

S. M. Sze M. K. Lee
Hsinchu, Taiwan Kaohsiung, Taiwan
August 2010



0

Why add a chapter 0 ??

Introduction

- ▶ 0.1 SEMICONDUCTOR DEVICES
 - ▶ 0.2 SEMICONDUCTOR TECHNOLOGY
 - ▶ SUMMARY
-

As an undergraduate in applied physics, electrical engineering, electronics engineering, or materials science, you might ask why you need to study semiconductor devices. The reason is that semiconductor devices are the foundation of the electronics industry, which is the largest industry in the world. A basic knowledge of semiconductor devices is essential to the understanding of advanced courses in electronics. This knowledge will also enable you to contribute to the Information Age, which is based on electronic technology.

Specifically, we cover the following topics:

- Four building blocks of semiconductor devices.
- Eighteen important semiconductor devices and their roles in electronic applications.
- Twenty three important semiconductor technologies and their roles in device processing.
- Technology trends toward high-density, high-speed, low-power consumption, and nonvolatility.

▶ 0.1 SEMICONDUCTOR DEVICES

Figure 1 shows the sales volume of the semiconductor-device-based electronics industry in the past 30 years and projects sales to the year 2020. Also shown are the gross world product (GWP) and the sales volumes of the automobile, steel, and semiconductor industries.^{1,2} We note that the electronics industry surpassed the automobile industry in 1998. If the current trends continue, in year 2020 the sales volume of the electronics industry will reach two trillion dollars and will constitute about 3% of GWP. It is expected that the electronic industry will remain the largest industry in the world throughout the 21st century. The semiconductor industry, which is a subset of the electronic industry, will surpass the steel industry around 2010 and constitute 25% of the electronics industry in 2020.

0.1.1 Device Building Blocks

Semiconductor devices have been studied for over 135 years.³ To date, there are 18 major devices, with over 140 device variations related to them.⁴ All these devices can be constructed from a small number of device building blocks.

Figure 2a is the metal-semiconductor interface, which is an intimate contact between a metal and a semiconductor. This building block was the first semiconductor device ever studied (in the year 1874). This interface can be used as a rectifying contact; that is, the device allows electrical current to flow easily only in one direction, or as an ohmic contact, which can pass current in either direction with a negligibly small voltage drop.

2 Semiconductors

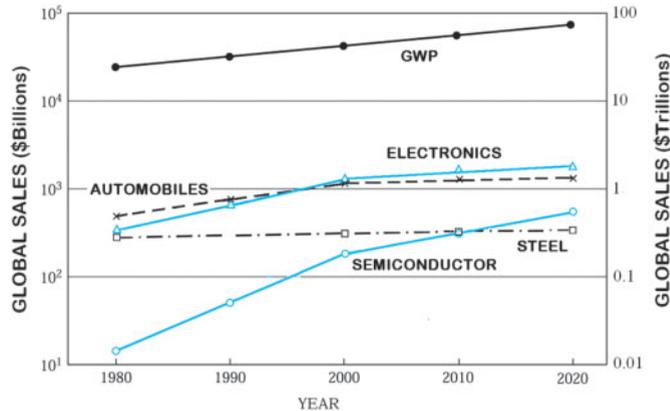


Fig. 1 Gross world product (GWP) and sales volumes of the electronics, automobile, semiconductor, and steel industries from 1980 to 2010 and projected to 2020.^{1,2}

We can use this interface to form many useful devices. For example, by using a rectifying contact as the *gate** and two ohmic contacts as the *source* and *drain*, we can form a *MESFET* (metal-semiconductor field-effect transistor), an important microwave device.

The second building block is the *p-n* junction (Fig. 2b), which is formed between a *p-type* (with positively charged carriers) and an *n-type* (with negatively charged carriers) semiconductor. The *p-n* junction is a key building block for most semiconductor devices, and *p-n* junction theory serves as the foundation of the physics of semiconductor devices. By combining two *p-n* junctions, that is, by adding another *p-type* semiconductor, we form the *p-n-p bipolar transistor*, which was invented in 1947 and had an unprecedented impact on the electronic industry. If we combine three *p-n* junctions to form a *p-n-p-n* structure, it becomes a switching device called a *thyristor*.

The third building block (Fig. 2c) is the heterojunction interface, that is, an interface formed between two dissimilar semiconductors. For example, we can use *gallium arsenide* (GaAs) and *aluminum arsenide* (AlAs) to form a heterojunction. Heterojunctions are the key components for high-speed and photonic devices.

Figure 2d shows the metal-oxide-semiconductor (MOS) structure. The structure can be considered a combination of a metal-oxide interface and an oxide-semiconductor interface. By using the MOS structure as the gate and two *p-n* junctions as the source and drain, we can form a *MOSFET* (MOS field-effect transistor). The MOSFET is the most important device for advanced *integrated circuits*, which contains tens of thousands of devices per integrated circuit chip.

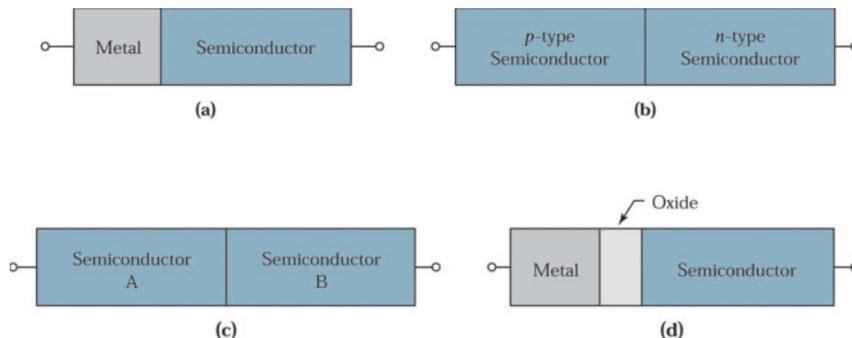


Fig. 2 Basic device building blocks. (a) Metal-semiconductor interface; (b) *p-n* junction; (c) heterojunction interface; and (d) metal-oxide-semiconductor structure.

*The italicized terms in this paragraph and in subsequent paragraphs are defined and explained in Part II of the book.

0.1.2 Major Semiconductor Devices

Some major semiconductor devices are listed in Table 1 in chronological order; those with a superscript *b* are two-terminal devices, and the others are three-terminal or four-terminal devices.³ The earliest systematic study of semiconductor devices (metal-semiconductor contacts) is generally attributed to Braun,⁵ who in 1874 discovered that the resistance of contacts between metals and metal sulfides (e.g., copper pyrite) depended on the magnitude and polarity of the applied voltage. The electroluminescence phenomenon (for the *light-emitting diode*) was discovered by Round⁶ in 1907. He observed the generation of yellowish light from a crystal of carborundum when he applied a potential of 10 V between two points on the crystal.

In 1947, the point-contact transistor was invented by Bardeen and Brattain.⁷ This was followed by Shockley's⁸ classic 1949 paper on *p-n* junctions and bipolar transistors. Figure 3 shows the first transistor. The two point contacts at the bottom of the triangular quartz crystal were made from two stripes of gold foil separated by about 50 μm ($1 \mu\text{m} = 10^{-4} \text{cm}$) and pressed onto a semiconductor surface. The semiconductor used was germanium. With one gold contact forward biased, that is, having positive voltage with respect to the third terminal, and the other reverse biased, the *transistor action* was observed: that is, the input signal was amplified. The bipolar transistor is a key semiconductor device and has ushered in the modern electronic era.

TABLE 1 Major Semiconductor Devices

| Year | Semiconductor Device ^a | Author(s)/Inventor(s) | Ref. |
|------|--|---------------------------------|-------|
| 1874 | Metal-semiconductor contact ^b | Braun | 5 |
| 1907 | Light emitting diode ^b | Round | 6 |
| 1947 | Bipolar transistor | Bardeen, Brattain, and Shockley | 7 |
| 1949 | <i>p-n</i> junction ^b | Shockley | 8 |
| 1952 | Thyristor | Ebers | 9 |
| 1954 | Solar cell ^b | Chapin, Fuller, and Pearson | 10 |
| 1957 | Heterojunction bipolar transistor | Kroemer | 11 |
| 1958 | Tunnel diode ^b | Esaki | 12 |
| 1960 | MOSFET | Kahng and Atalla | 13 |
| 1962 | Laser ^b | Hall et al. | 15 |
| 1963 | Heterostructure laser ^b | Kroemer, Alferov and Kazarinov | 16,17 |
| 1963 | Transferred-electron diode ^b | Gunn | 18 |
| 1965 | IMPATT diode ^b | Johnston, DeLoach, and Cohen | 19 |
| 1966 | MESFET | Mead | 20 |
| 1967 | Nonvolatile semiconductor memory | Kahng and Sze | 21 |
| 1970 | Charge-coupled device | Boyle and Smith | 23 |
| 1974 | Resonant tunneling diode ^b | Chang, Esaki, and Tsu | 24 |
| 1980 | MODFET | Mimura et al. | 25 |
| 2004 | 5 nm MOSFET | Yang et al. | 14 |

^aMOSFET, metal-oxide-semiconductor field-effect transistor; MESFET, metal-semiconductor field-effect transistor; MODFET, modulation-doped field-effect transistor.

^bDenotes a two-terminal device; others are a three- or four-terminal device.

4 Semiconductors

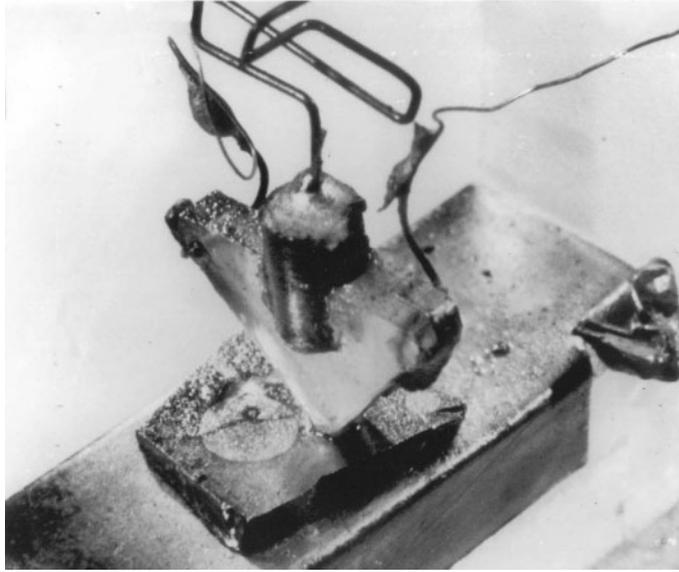


Fig. 3 The first transistor.⁷ (Photograph courtesy of Bell Laboratories, Alcatel-Lucent Co)

In 1952, Ebers⁹ developed the basic model for the thyristor, which is an extremely versatile switching device. The *solar cell* was developed by Chapin et al.¹⁰ in 1954 using a silicon $p-n$ junction. The solar cell is a major candidate for obtaining energy from the sun because it can convert sunlight directly to electricity and is environmentally benign. In 1957, Kroemer¹¹ proposed the heterojunction bipolar transistor to improve transistor performance; this device is potentially one of the fastest semiconductor devices. In 1958, Esaki¹² observed negative resistance characteristics in a heavily doped $p-n$ junction, which led to the discovery of the *tunnel diode*. The tunnel diode and the associated tunneling phenomenon are important for ohmic contacts and carrier transport through thin layers.

The most important device for advanced integrated circuits is the MOSFET, which was reported by Kahng and Atalla¹³ in 1960. Figure 4 shows the first device using a thermally oxidized silicon substrate. The device has a gate length of 20 μm and a gate oxide thickness of 100 nm (1 nm = 10^{-7} cm). The two keyholes are the source and drain contacts, and the top elongated area is the aluminum gate evaporated through a metal mask. Although present-day MOSFETs have been scaled down to the nanometer regime, the choice of silicon and thermally grown silicon dioxide used in the first MOSFET remains the most important combination of materials. The MOSFET and its related integrated circuits now constitute about 95% of the semiconductor device market. An ultrasmall MOSFET with a gate length of 5 nm has been demonstrated recently.¹⁴ This device can serve as the basis for the most advanced integrated circuit chips containing over one trillion ($>10^{12}$) devices.

In 1962, Hall et al.¹⁵ first achieved lasing in semiconductors, and in 1963, Kroemer¹⁶ and Alferov and Kazarinov¹⁷ proposed the *heterostructure laser*. These proposals laid the foundation for modern laser diodes, which can be operated continuously at room temperature. Laser diodes have revolutionized optoelectronic technology for a wide range of applications, including digital video disks, optical-fiber communication, laser printing, and atmospheric-pollution monitoring.

Three important microwave devices were invented or realized over the next three years. The first was the *transferred-electron diode* (TED; also called the Gunn diode) invented by Gunn¹⁸ in 1963. The TED is used extensively in such millimeter-wave applications as detection systems, remote controls, and microwave test instruments. The second device is the *IMPATT* diode; its operation was first observed by Johnston et al.¹⁹ in 1965. IMPATT diodes can generate the highest continuous wave (CW) power at millimeter-wave frequencies of all semiconductor devices. They are used in radar systems and alarm systems. The third device is the MESFET, invented by Mead²⁰ in 1966, which is a key device for monolithic microwave integrated circuits (MMIC).



Fig. 4 The first metal-oxide-semiconductor field-effect transistor.¹³ (Photograph courtesy of Bell Laboratories, Alcatel-Lucent Co)

Another important semiconductor memory device was invented by Kahng and Sze²¹ in 1967. This is the *nonvolatile semiconductor memory* (NVSM), which can retain its stored information for 10 to 100 years when the power supply is switched off. A schematic diagram of the first NVSM is shown in Fig. 5. Although it is similar to a conventional MOSFET, the major difference is the addition of the *floating gate*, in which semipermanent charge storage is possible. NVSM has revolutionized information-storage technology and enabled or enhanced the development of nearly all electronic products, especially portable electronic systems such as the cellular phone, digital camera, notebook computer, and global positioning system.²²

The *charge-coupled device* (CCD), invented by Boyle and Smith²³ in 1970, is used in digital cameras and optical sensing applications. The resonant tunneling diode (RTD) was first studied by Chang et al.²⁴ in 1974. RTD is the basis for most quantum-effect devices, which offer extremely high density, ultrahigh speed, and enhanced functionality because RTD significantly reduces the number of devices necessary to perform a given circuit function. In 1980, Minura et al.²⁵ developed the *MODFET* (modulation-doped field-effect transistor). With the proper selection of heterojunction materials, the MODFET is expected to be the fastest field-effect transistor.

Since the invention of the bipolar transistor in 1947, the number and variety of semiconductor devices have increased tremendously as advanced technology, new materials, and broadened comprehension have been applied to the creation of new devices. In Part II of the book, we consider all the devices listed in Table 1. It is hoped that this book can serve as a basis for understanding other devices not included here and perhaps not even conceived at the present time.

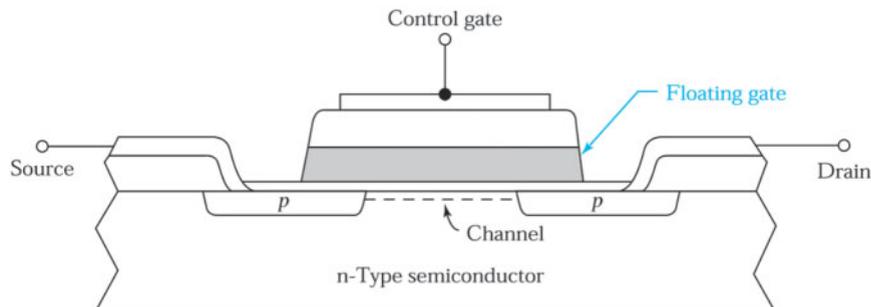


Fig. 5 A schematic diagram of the first nonvolatile semiconductor memory (NVSM) with a floating gate.²¹

► 0.2 SEMICONDUCTOR TECHNOLOGY

0.2.1 Key Semiconductor Technologies

Many important semiconductor technologies have been derived from processes invented centuries ago. For example, the lithography process was invented in 1798; in this first process, the pattern, or image, was transferred from a stone plate (lith- is Greek for ‘stone’).²⁶ In this section, we consider the milestones of technologies that were applied for the first time to semiconductor processing or were developed specifically for semiconductor-device fabrication.

Table 2 lists some key semiconductor technologies in chronological order. In 1918, Czochralski²⁷ developed a liquid-solid monocomponent growth technique. The Czochralski growth is the process used to grow most of the crystals from which silicon wafers are produced. Another growth technique, developed by Bridgman²⁸ in 1925, has been used extensively for growing gallium arsenide and related compound semiconductor crystals. Although the semiconductor properties of silicon have been widely studied since early 1940, the study of semiconductor compounds was neglected for a long time. In 1952, Welker²⁹ noted that gallium arsenide and its related III–V compounds were semiconductors. He was able to predict their characteristics and prove them experimentally. The technology and devices of these compounds have since been actively studied.

The diffusion of impurity atoms in semiconductors is important for device processing. Basic diffusion theory was considered by Fick³⁰ in 1855. The idea of using diffusion techniques to alter the type of conductivity in silicon was disclosed in a patent in 1952 by Pfann.³¹ The lithography process was first applied to semiconductor-device fabrication by Andrus in 1957.³² He used photosensitive etch-resistant polymers (photoresist) for pattern transfer. Lithography is a key technology for the semiconductor industry. The continued growth of the industry has been the direct result of improved lithographic technology. Lithography is also a significant economic factor, currently representing over 35% of the integrated-circuit manufacturing cost.

The oxide masking method was developed by Frosch and Derick³³ in 1957. They found that an oxide layer can prevent most impurity atoms from diffusing through it. In the same year, the epitaxial growth process based on chemical vapor deposition technique was developed by Sheftal et al.³⁴ Epitaxy (from the Greek epi, meaning on, and taxis, meaning arrangement) is a technique of crystal growth to form a thin layer of semiconductor materials on a crystal surface that has a lattice structure identical to that of the crystal. This method is important in improving device performance and creating novel device structures.

In 1958, Shockley³⁵ proposed a method of using ion implantation to dope semiconductors. Ion implantation has the capability of precisely controlling the number of implanted dopant atoms. Diffusion and ion implantation can complement each other for impurity doping. For example, diffusion can be used for high-temperature, deep-junction processes, whereas ion implantation can be used for lower-temperature, shallow-junction processes.

TABLE 2 KEY SEMICONDUCTOR TECHNOLOGIES

| Year | Technology ^a | Author(s)/Inventor(s) | Ref. |
|------|----------------------------|---------------------------------|------|
| 1918 | Czochralski crystal growth | Czochralski | 27 |
| 1925 | Bridgman crystal growth | Bridgman | 28 |
| 1952 | III-V compounds | Welker | 29 |
| 1952 | Diffusion | Pfann | 31 |
| 1957 | Lithographic photoresist | Andrus | 32 |
| 1957 | Oxide masking | Frosch and Derick | 33 |
| 1957 | Epitaxial CVD growth | Sheftal, Kokorish, and Krasilov | 34 |
| 1958 | Ion implantation | Shockley | 35 |
| 1959 | Hybrid integrated circuit | Kilby | 36 |

| | | | |
|------|-------------------------------|----------------------------|----|
| 1959 | Monolithic integrated circuit | Noyce | 37 |
| 1960 | Planar process | Hoerni | 38 |
| 1963 | CMOS | Wanlass and Sah | 39 |
| 1967 | DRAM | Dennard | 40 |
| 1969 | Polysilicon self-aligned gate | Kerwin, Klein, and Sarace | 41 |
| 1969 | MOCVD | Manasevit and Simpson | 42 |
| 1971 | Dry etching | Irving, Lemons, and Bobos | 43 |
| 1971 | Molecular beam epitaxy | Cho | 44 |
| 1971 | Microprocessor (4004) | Hoff et al. | 45 |
| 1981 | Atomic layer deposition | Suntola | 46 |
| 1982 | Trench isolation | Rung, Momose, and Nagakubo | 47 |
| 1989 | Chemical mechanical polishing | Davari et al. | 48 |
| 1993 | Copper interconnect | Paraszczak et al. | 49 |
| 2001 | 3D integration | Banerjee, et al. | 50 |
| 2003 | Immersion lithography | Owa, Nagasaka | 51 |

^aCVD, chemical vapor deposition; CMOS, complementary metal-oxide-semiconductor field-effect transistor; DRAM, dynamic random access memory; MOCVD, metalorganic CVD.

In 1959, a rudimentary integrated circuit (IC) was made by Kilby.³⁶ It contained one bipolar transistor, three resistors, and one capacitor, all made in germanium and connected by wire bonding—a hybrid circuit. Also in 1959, Noyce³⁷ created the monolithic IC by fabricating all devices in a single semiconductor substrate (monolith means ‘single stone’) and connecting the devices by aluminum metallization. Figure 6 shows the first monolithic IC of a flip-flop circuit containing six devices. The aluminum interconnection lines were obtained by etching evaporated aluminum layer over the entire oxide surface using the lithographic technique. These inventions laid the foundation for the rapid growth of the microelectronics industry.

The “planar” process was developed by Hoerni³⁸ in 1960. In this process, an oxide layer is formed on a semiconductor surface. With the help of a lithographic process, portions of the oxide can be removed and windows cut in the oxide. Impurity atoms will diffuse only through the exposed semiconductor surface, and $p-n$ junctions will form in the oxide window areas.

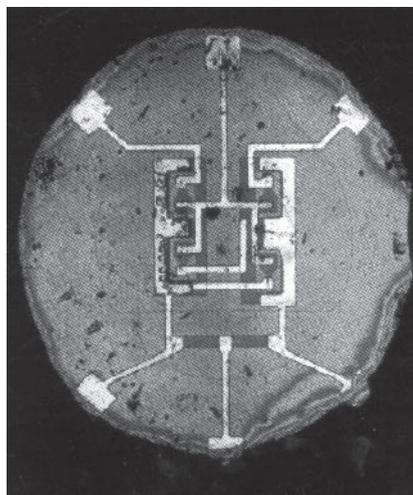


Fig. 6 The first monolithic integrated circuit.³⁷ (Photograph courtesy of Dr. G. Moore.)

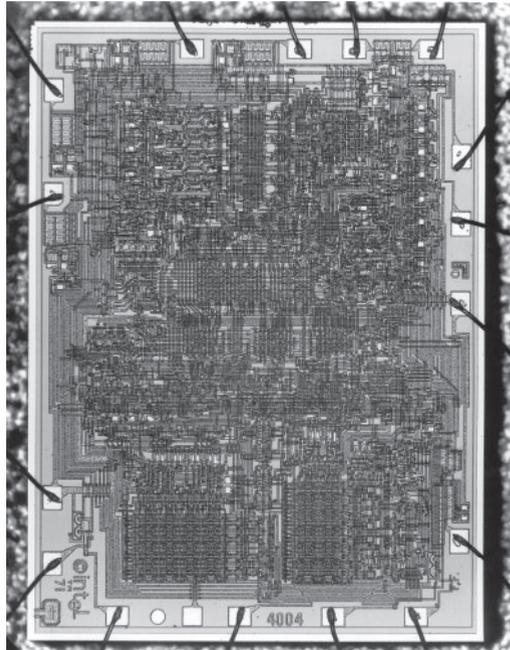


Fig. 7 The first microprocessor.⁴⁵ (Photograph courtesy of Intel Corp.)

As the complexity of the IC increased, we have moved from *NMOS* (*n*-channel MOSFET) to *CMOS* (complementary MOSFET) technology, which employs both NMOS and *PMOS* (*p*-channel MOSFET) to form the logic elements. The CMOS concept was proposed by Wanlass and Sah³⁹ in 1963. The advantage of CMOS technology is that logic elements draw significant current only during the transition from one state to another (e.g., from 0 to 1) and draw very little current between transitions, allowing power consumption to be minimized. CMOS technology is currently the dominant technology for advanced ICs.

In 1967, an important two-element circuit, the dynamic random-access memory (DRAM), was invented by Dennard.⁴⁰ The memory cell contains one MOSFET and one charge-storage capacitor. The MOSFET serves as a switch to charge or discharge the capacitor. Although DRAM is volatile and consumes relatively high power, we expect that DRAM will continue to be used in most electronic systems as an important working memory where information is held temporarily before being filed for long-term storage (e.g., in NVSM).

To improve the device performance, the polysilicon self-aligned gate process was proposed by Kerwin et al.⁴¹ in 1969. This process not only improved device reliability but also reduced parasitic capacitances. Also in 1969, the metalorganic chemical vapor deposition (MOCVD) method was developed by Manasevit and Simpson.⁴² This is a very important epitaxial growth technique for compound semiconductors such as GaAs.

As the device dimensions were reduced, a dry etching technique was developed to replace wet chemical etching for high-fidelity pattern transfer. This technique was initiated by Irving et al.⁴³ in 1971 using a $\text{CF}_4 - \text{O}_2$ gas mixture to etch silicon wafers. Another important technique developed in the same year by Cho is molecular beam epitaxy.⁴⁴ This technique has the advantage of near-perfect vertical control of composition and doping down to atomic dimensions. It is responsible for the creation of numerous photonic devices and quantum-effect devices.

In 1971, the first microprocessor was made by Hoff et al.⁴⁵ They put the entire central processing unit (CPU) of a simple computer on one chip. This was a four-bit microprocessor (Intel 4004), shown in Fig. 7, with a chip size of $3 \text{ mm} \times 4 \text{ mm}$, and it contained 2300 MOSFETs and operated at 0.1 MIPS (million instructions per second). It was fabricated by a *p*-channel, polysilicon gate process using an $8 \mu\text{m}$ design rule. This microprocessor performed as well as those in \$300,000 IBM computers of the early 1960s—each of which needed a CPU the size of a large desk. This was a major breakthrough for the semiconductor industry. Currently, microprocessors constitute the largest segment of the industry.

Since early 1980, many new technologies have been developed to meet the requirements of ever-shrinking minimum feature lengths. An important technique, atomic layer deposition (ALD), was developed for nanoscaled dielectric film deposition by Suntola in 1981.⁴⁶ This deposition technique involves exposing the chemical precursors to the growth surface in a sequential, one-at-a-time manner. The film thickness can be reliably controlled down to atomic dimensions.

The trench isolation technology was introduced by Rung et al.⁴⁷ in 1982 to isolate CMOS devices and has ultimately replaced all other isolation methods. In 1989, Davari et al.⁴⁸ developed the chemical-mechanical polishing method for global planarization of the interlayer dielectrics. This is a key process in multilevel metallization.

At submicron dimensions, a widely known failure mechanism is electromigration, which is the transport of metal ions through a conductor due to the passage of an electrical current. Although aluminum has been used since the early 1960s as the interconnect material, it suffers from electromigration at high electrical current. The copper interconnect was introduced in 1993 by Paraszcak et al.⁴⁹ to replace aluminum for minimum feature lengths below 100 nm.

The increased component density and improved fabrication technology have helped realize the system-on-a-chip (SOC), which is an IC chip containing a complete electronic system. The system-on-a-chip was integrated into a three-dimensional (3-D) system with improved performance by Banerjee in 2001.⁵⁰

In order to extend the optical photolithography to the nanoscale regime, Owa et al. in 2003⁵¹ developed immersion lithography through the addition of water between the exposure lens and the wafer surface. Immersion lithography increases the resolution by a factor equal to the refractive index of the liquid, and the minimum feature size can be made below 45 nm. In Part III of this book, we consider all the technologies listed in Table 2.

0.2.2 Technology Trends

Since the beginning of the microelectronics era, the smallest line width or the minimum feature length of an integrated circuit has been reduced at a rate of about 13% per year.⁵² At that rate, the minimum feature length will shrink to about 10 nm in the year 2020. Figure 8 shows the minimum feature length versus year of first production from 1978 to 2010 and projected to 2020. In 2002 we entered the nanoelectronics era by reducing the minimum feature length below 100 nm.

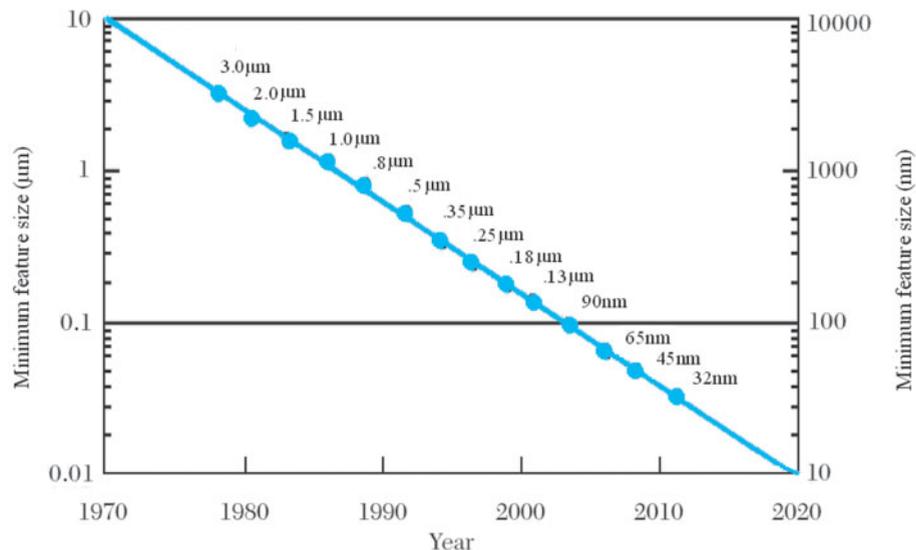


Fig. 8 Exponential decrease of the minimum feature length versus time.⁵²

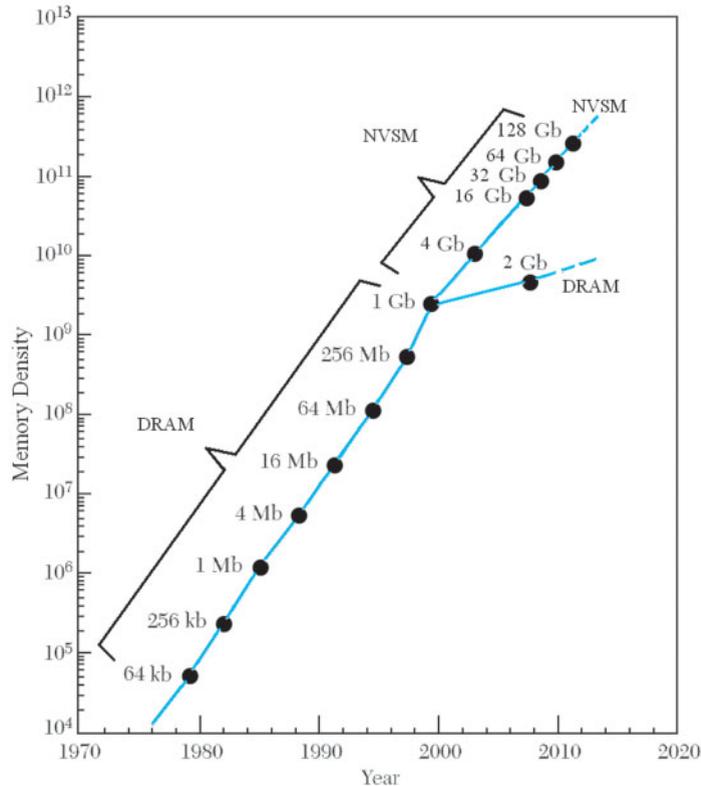


Fig. 9 Exponential increase in dynamic random access memory (DRAM) density and nonvolatile semiconductor memory (NVSM) density versus year.⁵²

Device miniaturization results in reduced unit cost per circuit function. For example, the cost per bit of memory chips has halved every two years for successive generations of DRAMs. As device dimension decreases, the intrinsic switching time also decreases: device speed has improved by five orders of magnitude since 1959. Higher speeds lead to expanded IC functional throughput rates. In the future, digital ICs will be able to perform data processing and numerical computation at terabit-per-second rates. As devices become smaller, they consume less power. Therefore, device miniaturization also reduces the energy used for each switching operation. The energy dissipated per logic gate has decreased by over ten million times since 1959.

Figure 9 shows the exponential increase in the actual memory density versus year of first production over the past 30 years. We note that the DRAM density increased by a factor of 2 every 18 months from 1978 to 2000. After 2000, the growth rate of DRAM density slowed down considerably. On the other hand, NVSM density has continued the original growth rate of DRAM density, i.e., doubling every 18 months. If the trends continue, we expect that NVSM density will increase to 1000 Gb or 1 terabits (10^{12} bits) around 2015.

Figure 10 shows the exponential increase of microprocessor computational power. The computational power also increases approximately by a factor of 2 every 18 months. Currently, a Pentium-based personal computer has greater computational power than the CRAY 1 supercomputer of the late 1960s; yet, the PC is four orders of magnitude smaller. If the trends continue, we will reach 10^7 MIP (million instructions per second) around 2015.

Figure 11 illustrates growth curves for different technology drivers.⁵³ At the beginning of the modern electronic era (1950–1970), the bipolar transistor was the technology driver. From 1970 to 1990, the DRAM and the microprocessor based on MOS devices were the technology drivers because of the rapid growth of personal computers and advanced electronic systems. Since 1990, nonvolatile semiconductor memory has been the technology driver, mainly because of the rapid growth of portable electronic systems.

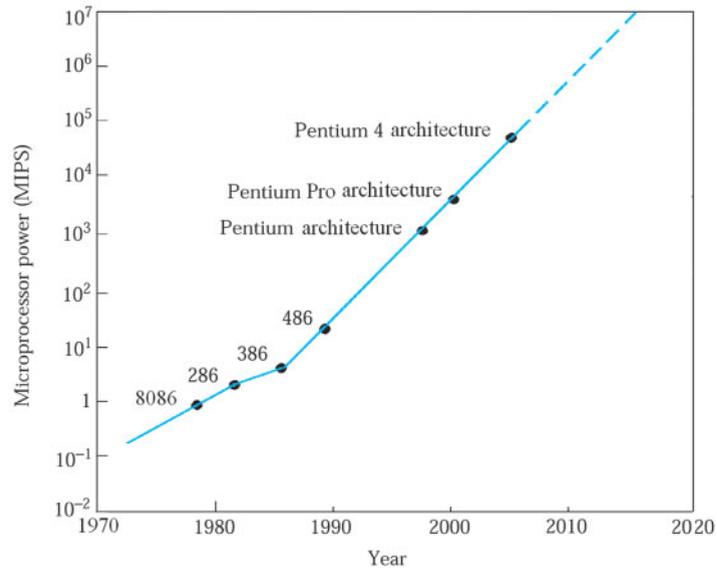


Fig. 10 Exponential increase in microprocessor computational power versus year. (From Intel Corp.)

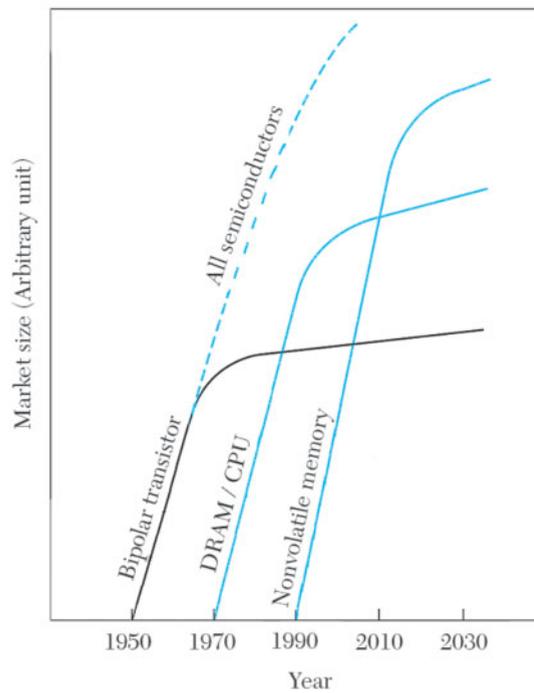


Fig. 11 Growth curves for different technology drivers.⁵³

12 Semiconductors

► SUMMARY

Although the field of *semiconductor devices* is a relatively new area of study,* it has had enormous impact on our society and the global economy. This is because semiconductor devices are the foundation of the largest industry in the world—the electronics industry.

This introductory chapter has presented a historical review of major semiconductor devices from the first study of metal-semiconductor contact in 1874 to the fabrication of an ultrasmall 5-nm MOSFET in 2004. Of particular importance are the invention of the bipolar transistor in 1947, which ushered in the modern electronic era; the development in 1960 of the MOSFET, the most important device for integrated circuits; and the invention of the nonvolatile semiconductor memory in 1967, which has been the technology driver of the electronic industry since 1990.

We have also described key semiconductor technologies. The origins of many technologies can be traced back to the late eighteenth and early nineteenth centuries. Of particular importance are the development of the lithographic photoresist in 1957, which established the basic pattern-transfer process for semiconductor devices; the invention of the integrated circuits in 1959, which was seminal to the rapid growth of the microelectronic industry; and the developments of the DRAM in 1967 and the microprocessor in 1971, which constitute the two largest segments of the semiconductor industry.

There is a vast literature on semiconductor-device physics and technology.⁵⁴ To date, more than 500,000 papers have been published in this field. In this book, each chapter deals with a major device or a key technology. Each is presented in a clear and coherent fashion without heavy reliance on the original literature. However, we have selected a few important papers at the end of each chapter for reference and for further reading.

► REFERENCES

1. *2009 Semiconductor Industry Report*, Ind. Technol. Res. Inst., Hsinchu, Taiwan, 2009.
2. Data from IC Insights, 2009.
3. Most of the classic device papers are collected in S. M. Sze, Ed., *Semiconductor Devices: Pioneering Papers*, World Sci., Singapore, 1991.
4. K. K. Ng, *Complete Guide to Semiconductor Devices*, 2nd Ed., Wiley Interscience, New York, 2002.
5. F. Braun, “Über die Stromleitung durch Schwefelmetalle,” *Ann. Phys. Chem.*, **153**, 556 (1874).
6. H. J. Round, “A Note On Carborundum,” *Electron. World*, **19**, 309 (1907).
7. J. Bardeen and W. H. Brattain, “The Transistor, a Semiconductor Triode,” *Phys. Rev.*, **71**, 230 (1948).
8. W. Shockley, “The Theory of $p-n$ Junction in Semiconductors and $p-n$ Junction Transistors,” *Bell Syst. Tech. J.*, **28**, 435 (1949).
9. J. J. Ebers, “Four Terminal $p-n-p-n$ Transistors,” *Proc. IRE*, **40**, 1361 (1952).
10. D. M. Chapin, C. S. Fuller, and G. L. Pearson, “A New Silicon $p-n$ Junction Photocell for Converting Solar Radiation into Electrical Power,” *J. Appl. Phys.*, **25**, 676 (1954).
11. H. Kroemer, “Theory of a Wide-Gap Emitter for Transistors,” *Proc. IRE*, **45**, 1535 (1957).
12. L. Esaki, “New Phenomenon in Narrow Germanium $p-n$ Junctions,” *Phys. Rev.*, **109**, 603 (1958).

*Semiconductor devices and materials have been studied since the early nineteenth century. However, many traditional devices and materials have been studied for a much longer time. For example, steel was first studied in 1200 BC, over 3000 years ago.

13. D. Kahng and M. M. Atalla, "Silicon-Silicon Dioxide Surface Device," in *IRE Device Research Conference*, Pittsburgh, 1960. (The paper can be found in Ref. 3.)
14. F. L. Yang et al., "5 nm Gate Nanowire FinFET", *Symp. VLSI Tech.*, June 15, 2004.
15. R. N. Hall et al., "Coherent Light Emission from GaAs Junctions," *Phys. Rev. Lett.*, **9**, 366 (1962).
16. H. Kroemer, "A Proposed Class of Heterojunction Injection Lasers," *Proc. IEEE*, **51**, 1782 (1963).
17. I. Alferov and R. F. Kazarinov, "Semiconductor Laser with Electrical Pumping," U.S.S.R. Patent 181, 737 (1963).
18. J. B. Gunn, "Microwave Oscillations of Current in III-V Semiconductors," *Solid State Commun.*, **1**, 88 (1963).
19. R. L. Johnston, B. C. DeLoach, Jr., and B. G. Cohen, "A Silicon Diode Microwave Oscillator," *Bell Syst. Tech. J.*, **44**, 369 (1965).
20. C. A. Mead, "Schottky Barrier Gate Field Effect Transistor," *Proc. IEEE*, **54**, 307 (1966).
21. D. Kahng and S. M. Sze, "A Floating Gate and Its Application to Memory Devices," *Bell Syst. Tech. J.*, **46**, 1288 (1967).
22. C. Y. Lu and H. Kuan, "Nonvolatile Semiconductor Memory Revolutionizing Information Storage", *IEEE Nanotechnology Mag.*, **3**, 4, (2009).
23. W. S. Boyle and G. E. Smith, "Charge Coupled Semiconductor Devices," *Bell Syst. Tech. J.*, **49**, 587 (1970).
24. L. L. Chang, L. Esaki, and R. Tsu, "Resonant Tunneling in Semiconductor Double Barriers," *Appl. Phys. Lett.*, **24**, 593 (1974).
25. T. Mimura, et al., "A New Field-Effect Transistor with Selectively Doped GaAs/n-Al_xGa_{1-x} as Heterojunction," *Jpn. J. Appl. Phys.*, **19**, L225 (1980).
26. M. Hefner, "The Photoresist Story," *J. Photo. Sci.*, **12**, 181 (1964).
27. J. Czochralski, "Ein neues Verfahren zur Messung der Kristallisationsgeschwindigkeit der Metalle," *Z. Phys. Chem.*, **92**, 219 (1918).
28. P. W. Bridgman, "Certain Physical Properties of Single Crystals of Tungsten, Antimony, Bismuth, Tellurium, Cadmium, Zinc, and Tin," *Proc. Am. Acad. Arts Sci.*, **60**, 303 (1925).
29. H. Welker, "Über Neue Halbleitende Verbindungen," *Z. Naturforsch.*, **7a**, 744 (1952).
30. A. Fick, "Ueber Diffusion," *Ann. Phys. Lpz.*, **170**, 59 (1855).
31. W. G. Pfann, "Semiconductor Signal Translating Device," U.S. Patent 2, 597,028 (1952).
32. J. Andrus, "Fabrication of Semiconductor Devices," U.S. Patent 3,122,817 (filed 1957; granted 1964).
33. C. J. Frosch and L. Derick, "Surface Protection and Selective Masking During Diffusion in Silicon," *J. Electrochem. Soc.*, **104**, 547 (1957).
34. N. N. Sheftal, N. P. Kokorish, and A. V. Krasilov, "Growth of Single-Crystal Layers of Silicon and Germanium from the Vapor Phase," *Bull. Acad. Sci U.S.S.R., Phys. Ser.*, **21**, 140 (1957).
35. W. Shockley, "Forming Semiconductor Device by Ionic Bombardment," U.S. Patent 2,787,564 (1958).
36. J. S. Kilby, "Invention of the Integrated Circuit," *IEEE Trans. Electron Devices*, **ED-23**, 648 (1976), U.S. Patent 3,138,743 (filed 1959, granted 1964).

14 Semiconductors

37. R. N. Noyce, "Semiconductor Device-and-Lead Structure," U.S. Patent 2,981,877 (filed 1959, granted 1961).
38. J. A. Hoerni, "Planar Silicon Transistors and Diodes," *IRE Int. Electron Devices Meet.*, Washington D.C. (1960).
39. F. M. Wanlass and C. T. Sah, "Nanowatt Logics Using Field-Effect Metal-Oxide Semiconductor Triodes," *Tech. Dig. IEEE Int. Solid-State Circuit Conf.*, p. 32 (1963).
40. R. M. Dennard, "Field Effect Transistor Memory," U.S. Patent 3,387,286 (filed 1967, granted 1968).
41. R. E. Kerwin, D. L. Klein, and J. C. Sarace, "Method for Making MIS Structure," U.S. Patent 3,475,234 (1969).
42. H. M. Manasevit and W. I. Simpson, "The Use of Metal–Organic in the Preparation of Semiconductor Materials. I. Epitaxial Gallium-V Compounds," *J. Electrochem. Soc.*, **116**, 1725 (1969).
43. S. M. Irving, K. E. Lemons, and G. E. Bobos, "Gas Plasma Vapor Etching Process," U.S. Patent 3,615,956 (1971).
44. A. Y. Cho, "Film Deposition by Molecular Beam Technique," *J. Vac. Sci. Technol.*, **8**, S 31 (1971).
45. The inventors of the microprocessor are M. E. Hoff, F. Faggin, S. Mazor, and M. Shima. For a profile of M. E. Hoff, see *Portraits in Silicon* by R. Slater, p. 175, MIT Press, Cambridge, 1987.
46. T. Suntola, "Atomic Layer Epitaxy", *Tech. Digest of ICVGE-5*, San Diego, 1981.
47. R. Rung, H. Momose, and Y. Nagakubo, "Deep Trench Isolated CMOS Devices," *Tech. Dig. IEEE Int. Electron Devices Meet.*, p. 237 (1982).
48. B. Davari et al., "A New Planarization Technique, Using a Combination of RIE and Chemical Mechanical Polish (CMP)," *Tech. Dig. IEEE Int. Electron Devices Meet.*, p. 61 (1989).
49. J. Paraszczak et al., "High Performance Dielectrics and Processes for ULSI Interconnection Technologies," *Tech. Dig. IEEE Int. Electron Devices Meet.*, p.261 (1993).
50. K. Banerjee et al., "3-D ICs: A Novel Chip Design for Improving Deep-Submicrometer Interconnect Performance and Systems-on-Chip Integration," *Proc. IEEE*, **89**, 602 (2001)
51. S. Owa and H. Nagasaka, "Immersion Lithography; Its Potential Performance and Issues," *Proc. SPIE*, 5040, 724-33, (2003).
52. *The International Technology Roadmap for Semiconductor*, Semiconductor Ind. Assoc., San Jose, 2010.
53. F. Masuoka, "Flash Memory Technology," *Proc. Int. Electron Devices Mater. Symp.*, 83, Hsinchu, Taiwan (1996).
54. From *INSPEC* database, National Chaio Tung University, Hsinchu, Taiwan, 2010.

Energy Bands and Carrier Concentration in Thermal Equilibrium

- ▶ 1.1 SEMICONDUCTOR MATERIALS
 - ▶ 1.2 BASIC CRYSTAL STRUCTURES
 - ▶ 1.3 VALENCE BONDS
 - ▶ 1.4 ENERGY BANDS
 - ▶ 1.5 INTRINSIC CARRIER CONCENTRATION
 - ▶ 1.6 DONORS AND ACCEPTORS
 - ▶ SUMMARY
-

In this chapter, we consider some basic properties of semiconductors. We begin with a discussion of crystal structure, which is the arrangement of atoms in a solid. We then present the concepts of valence bonds and energy bands, which relate to conduction in semiconductors. Finally, we discuss the concept of carrier concentration in thermal equilibrium. These concepts are used throughout this book.

Specifically, we cover the following topics:

- Element and compound semiconductors and their basic properties.
- The diamond structure and its related crystal planes.
- The bandgap and its impact on electrical conductivity.
- The intrinsic carrier concentration and its dependence on temperature.
- The Fermi level and its dependence on carrier concentration.

▶ 1.1 SEMICONDUCTOR MATERIALS

Solid-state materials can be grouped into three classes—insulators, semiconductors, and conductors. Figure 1 shows the range of electrical conductivities σ (and the corresponding resistivities $\rho = 1/\sigma$)^{*} associated with some important materials in each of the three classes. Insulators such as fused quartz and glass have very low conductivities, on the order of 10^{-18} – 10^{-8} S/cm; and conductors such as aluminum and silver have high conductivities, typically from 10^4 to 10^6 S/cm.[§] Semiconductors have conductivities between those of

^{*}A list of symbols is given in Appendix A.

[§]The international system of units is presented in Appendix B.

16 Semiconductors

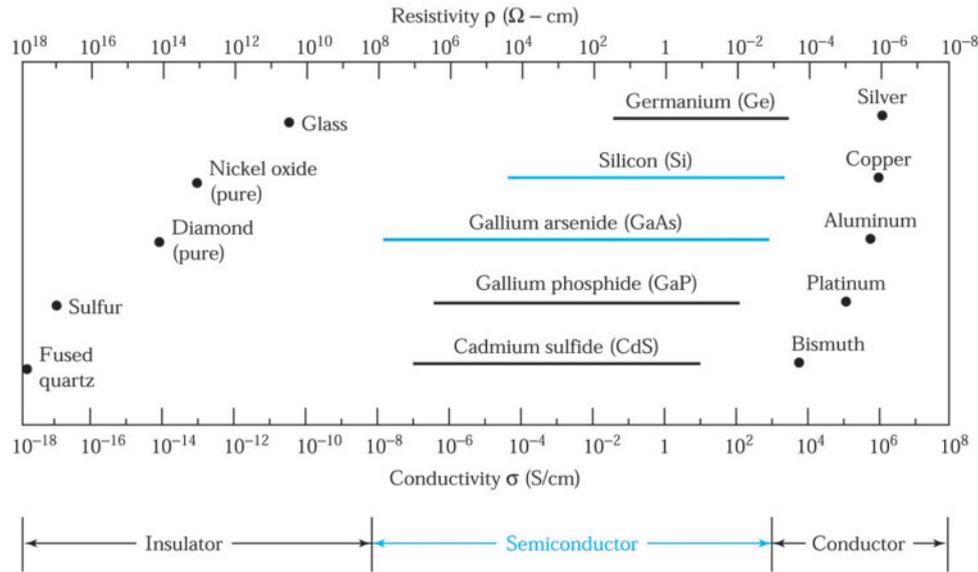


Fig. 1 Typical range of conductivities for insulators, semiconductors, and conductors.

insulators and those of conductors. The conductivity of a semiconductor is generally sensitive to temperature, illumination, magnetic field, and minute amounts of impurity atoms (typically, about $1\ \mu\text{g}$ to $1\ \text{g}$ of impurity atoms in $1\ \text{kg}$ of semiconductor materials). This sensitivity in conductivity makes the semiconductor one of the most important materials for electronic applications.

1.1.1 Element Semiconductors

The study of semiconductor materials began in the early nineteenth century.¹ Over the years many semiconductors have been investigated. Table 1 shows a portion of the periodic table related to semiconductors. The element semiconductors, those composed of single species of atoms, such as silicon (Si) and germanium (Ge), can be found in Column IV. In the early 1950s, germanium was the major semiconductor material. Since the early 1960s silicon has become a practical substitute and has now virtually supplanted germanium as a semiconductor material. The main reasons we now use silicon are that silicon devices exhibit better properties at room temperature, and high-quality silicon dioxide can be grown thermally. There are also economic considerations. Device-grade silicon costs much less than any other semiconductor material. Silicon in the form of silica and silicates comprises 25% of the Earth's crust, and silicon is second only to oxygen in abundance. Currently, silicon is one of the most studied elements in the periodic table, and silicon technology is by far the most advanced among all semiconductor technologies.

TABLE 1 Portion of the Periodic Table Related to Semiconductors

| Period | Column II | III | IV | V | VI |
|--------|-----------|----------|-----------|------------|-----------|
| 2 | | B | C | N | O |
| | | Boron | Carbon | Nitrogen | Oxygen |
| 3 | Mg | Al | Si | P | S |
| | Magnesium | Aluminum | Silicon | Phosphorus | Sulfur |
| 4 | Zn | Ga | Ge | As | Se |
| | Zinc | Gallium | Germanium | Arsenic | Selenium |
| 5 | Cd | In | Sn | Sb | Te |
| | Cadmium | Indium | Tin | Antimony | Tellurium |
| 6 | Hg | | Pb | | |
| | Mercury | | Lead | | |

1.1.2 Compound Semiconductors

In recent years a number of compound semiconductors have found applications for various devices. The important compound semiconductors as well as the two-element-semiconductors are listed² in Table 2. A binary compound semiconductor is a combination of two elements from the periodic table. For example, gallium arsenide (GaAs) is a III-V compound that is a combination of gallium (Ga) from Column III and arsenic (As) from Column V.

In addition to binary compounds, ternary compounds and quaternary compounds are made for special applications. The alloy semiconductor $\text{Al}_x\text{Ga}_{1-x}\text{As}$, which has Al and Ga from Column III and As from Column V is an example of a ternary compound, whereas quaternary compounds of the form $\text{A}_x\text{B}_{1-x}\text{C}_y\text{D}_{1-y}$ can be obtained from the combination of many binary and ternary compound semiconductors. For example, GaP, InP, InAs, and GaAs can be combined to yield the alloy semiconductor $\text{Ga}_x\text{In}_{1-x}\text{As}_y\text{P}_{1-y}$. Compared with the element semiconductors, the preparation of compound semiconductors in single-crystal form usually involves much more complex processes.

Many of the compound semiconductors have electrical and optical properties that are different from those of silicon. These semiconductors, especially GaAs, are used mainly for high-speed electronic and photonic applications. Although we do not know as much about the technology of compound semiconductors as we do about that of silicon, advances in silicon technology have also helped progress in compound semiconductor technology. In this book we are concerned mainly with device physics and processing technology of silicon and gallium arsenide. A detailed discussion of the crystal growth of silicon and gallium arsenide can be found in Chapter 11.

► 1.2 BASIC CRYSTAL STRUCTURES

The semiconductor materials we will be studying are single crystals; that is, the atoms are arranged in a three-dimensional periodic fashion. The periodic arrangement of atoms in a crystal is called a *lattice*. In a crystal, an atom never strays far from a single, fixed position. The thermal vibrations associated with the atom are centered about this position. For a given semiconductor, there is a *unit cell* that is representative of the entire lattice; by repeating the unit cell throughout the crystal, one can generate the entire lattice.

TABLE 2 Semiconductor Materials²

| General Classification | Semiconductor | | |
|------------------------|----------------------------|-------------------------------------|---------------------|
| | Symbol | Name | |
| Element | Si | Silicon | |
| | Ge | Germanium | |
| Binary compound | IV-IV ----- | SiC | Silicon carbide |
| | III-V ----- | AlP | Aluminum phosphide |
| | | AlAs | Aluminum arsenide |
| | | AlSb | Aluminum antimonide |
| | | GaN | Gallium nitride |
| | | GaP | Gallium phosphide |
| | | GaAs | Gallium arsenide |
| | | GaSb | Gallium antimonide |
| | | InP | Indium phosphide |
| | | InAs | Indium arsenide |
| | | InSb | Indium antimonide |
| | II-VI ----- | ZnO | Zinc oxide |
| | | ZnS | Zinc sulfide |
| | | ZnSe | Zinc selenide |
| | | ZnTe | Zinc telluride |
| | | CdS | Cadmium sulfide |
| | | CdSe | Cadmium selenide |
| | | CdTe | Cadmium telluride |
| | | HgS | Mercury sulfide |
| IV-VI ----- | PbS | Lead sulfide | |
| | PbSe | Lead selenide | |
| | PbTe | Lead telluride | |
| | | | |
| Ternary compound | $Al_xGa_{1-x}As$ | Aluminum gallium arsenide | |
| | $Al_xIn_{1-x}As$ | Aluminum indium arsenide | |
| | $GaAs_{1-x}P_x$ | Gallium arsenic phosphide | |
| | $Ga_xIn_{1-x}N$ | Gallium indium nitride | |
| | $Ga_xIn_{1-x}As$ | Gallium indium arsenide | |
| Quaternary compound | $Ga_xIn_{1-x}P$ | Gallium indium phosphide | |
| | $Al_xGa_{1-x}As_ySb_{1-y}$ | Aluminum gallium arsenic antimonide | |
| | $Ga_xIn_{1-x}As_{1-y}P_y$ | Gallium indium arsenic phosphide | |

1.2.1 Unit Cell

A generalized primitive three-dimensional unit cell is shown in Fig. 2. The relationship between this cell and the lattice is characterized by three vectors \mathbf{a} , \mathbf{b} , and \mathbf{c} , which need not be perpendicular to each other and may or may not be equal in length. Every equivalent lattice point in the three-dimensional crystal can be found using the set

$$\mathbf{R} = m\mathbf{a} + n\mathbf{b} + p\mathbf{c}, \quad (1)$$

where m , n , and p are integers.

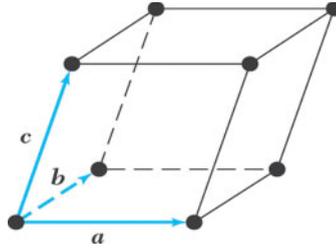


Fig. 2 A generalized primitive unit cell.

Figure 3 shows some basic cubic-crystal unit cells. Figure 3a shows a simple cubic (sc) crystal; it has an atom at each corner of the cubic lattice, and each atom has six equidistant nearest-neighbor atoms. The dimension a is called the lattice constant. In the periodic table, only polonium is crystallized in the simple cubic lattice. Figure 3b is a body-centered cubic (bcc) crystal where, in addition to the eight corner atoms, an atom is located at the center of the cube. In a bcc lattice, each atom has eight nearest-neighbor atoms. Crystals exhibiting bcc lattices include those of sodium and tungsten. Figure 3c shows the face-centered cubic (fcc) crystal that has one atom at each of the six cubic faces in addition to the eight corner atoms. In this case, each atom has 12 nearest-neighbor atoms. A large number of elements exhibit the fcc lattice form, including aluminum, copper, gold, and platinum.

► EXAMPLE 1

If we pack hard spheres in a bcc lattice so that the atom in the center just touches the atoms at the corners of the cube, find the fraction of the bcc unit cell volume filled with hard spheres.

SOLUTION Each corner sphere in a bcc unit cell is shared with eight neighboring cells; thus, each unit cell contains one-eighth of a sphere at each of the eight corners for a total of one sphere. In addition, each unit cell contains one central sphere. We have the following:

$$\text{Spheres (atoms) per unit cell} = (1/8) \times 8 \text{ (corner)} + 1 \text{ (center)} = 2;$$

$$\text{Nearest-neighbor distance (along the diagonal AE in Fig. 3b)} = a\sqrt{3}/2;$$

$$\text{Radius of each sphere} = a\sqrt{3}/4;$$

$$\text{Volume of each sphere} = 4\pi/3 \times (a\sqrt{3}/4)^3 = \pi a^3 \sqrt{3}/16; \text{ and}$$

$$\text{Maximum fraction of unit cell filled} = \text{Number of spheres} \times \text{volume of each sphere} / \text{total volume of each unit cell} = 2(\pi a^3 \sqrt{3}/16)/a^3 = \pi \sqrt{3}/8 \approx 0.68.$$

Therefore, about 68% of the bcc unit cell volume is filled with hard spheres, and about 32% of the volume is empty. ◀

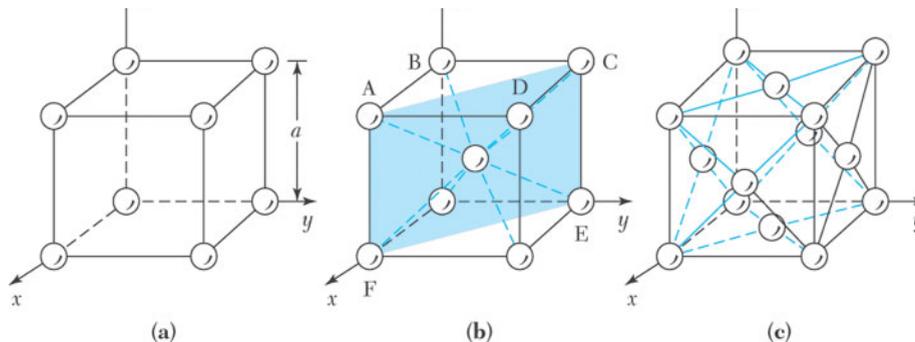


Fig. 3 Three cubic-crystal unit cells. (a) Simple cubic. (b) Body-centered cubic. (c) Face-centered cubic.

1.2.2 The Diamond Structure

The element semiconductors, silicon and germanium, have the diamond lattice structure shown in Fig. 4a. This structure also belongs to the fcc crystal family and can be seen as two interpenetrating fcc sublattices with one sublattice displaced from the other by one-quarter of the distance along the body diagonal of the cube (i.e., a displacement of $a\sqrt{3}/4$). Although chemically identical, the two sets of atoms belonging to the two sublattices are different in terms of the crystal structure. It can be seen in Fig. 4a that if a corner atom has one nearest neighbor in the body diagonal direction, then it has no nearest neighbor in the reverse direction. Consequently, two such atoms are required in the unit cell. Alternatively, a unit cell of a diamond lattice consists of a tetrahedron in which each atom is surrounded by four equidistant nearest neighbors that lie at the corners (the spheres connected by darkened bars in Fig. 4a).

Most of the III-V compound semiconductors (e.g., GaAs) have a *zincblende lattice*, shown in Fig. 4b, which is identical to a diamond lattice except that one fcc sublattice has Column III atoms (Ga) and the other has Column V atoms (As). Appendix F gives a summary of the lattice constants and other properties of important element and binary compound semiconductors.

► EXAMPLE 2

At 300 K the lattice constant for silicon is 5.43 \AA . Calculate the number of silicon atoms per cubic centimeter and the density of silicon at room temperature.

SOLUTION There are eight atoms per unit cell. Therefore,

$$8/a^3 = 8/(5.43 \times 10^{-8})^3 = 5 \times 10^{22} \text{ atoms/cm}^3; \text{ and}$$

$$\text{Density} = \text{no. of atoms/cm}^3 \times \text{atomic weight/Avogadro's number} = 5 \times 10^{22} (\text{atoms/cm}^3) \times 28.09 (\text{g/mol})/6.02 \times 10^{23} (\text{atoms/mol}) = 2.33 \text{ g/cm}^3. \quad \blacktriangleleft$$

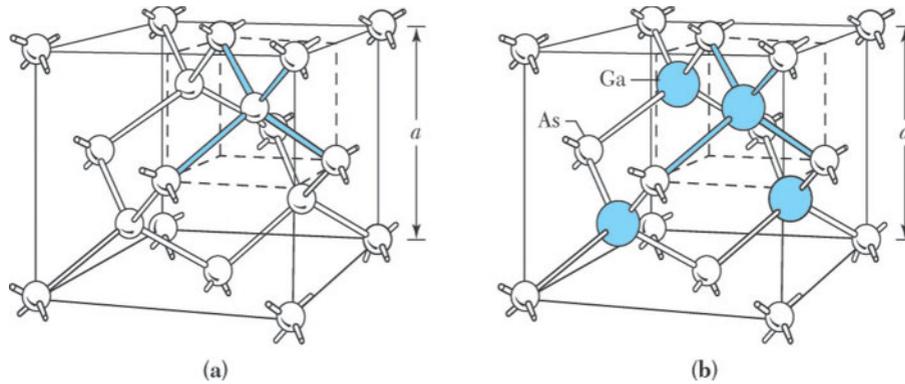


Fig. 4 (a) Diamond lattice. (b) Zincblende lattice.

1.2.3 Crystal Planes and Miller Indices

In Fig. 3b we note that there are four atoms in the $ABCD$ plane and five atoms in the $ACEF$ plane (four atoms from the corners and one from the center) and that the atomic spacing is different in the two planes. Therefore, the crystal properties along different planes are different, and the electrical and other device characteristics can be dependent on the crystal orientation. A convenient method of defining the various planes in a crystal is to use *Miller indices*.³ These indices are obtained using the following steps:

1. Find the intercepts of the plane on the three Cartesian coordinates in terms of the lattice constant.
2. Take the reciprocals of these numbers and reduce them to the smallest three integers having the same ratio.
3. Enclose the result in parentheses (hkl) as the Miller indices for a single plane.

► EXAMPLE 3

As shown in Fig. 5, the plane has intercepts at a , $3a$, and $2a$ along the three coordinates. Taking the reciprocals of these intercepts, we get 1 , $1/3$, and $1/2$. The smallest three integers having the same ratio are 6 , 2 , and 3 (obtained by multiplying each fraction by 6). Thus, the plane is referred to as a (623) -plane. ◀

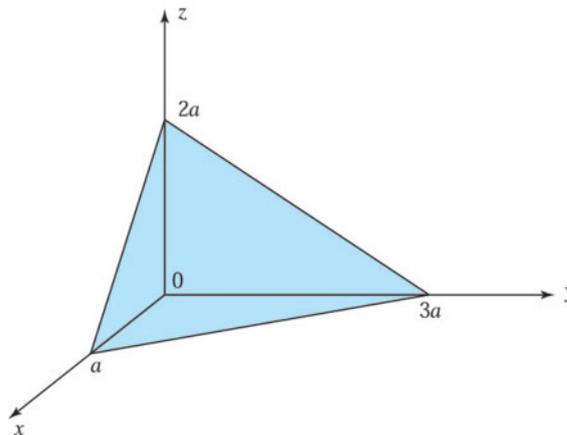


Fig. 5 A (623) -crystal plane.

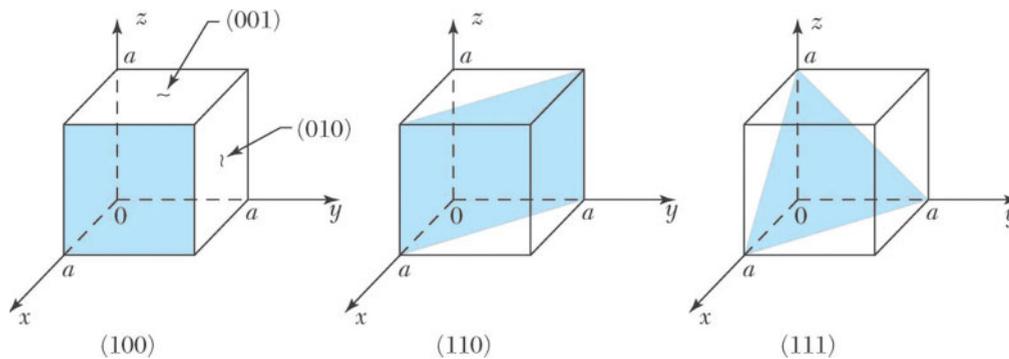


Fig. 6 Miller indices of some important planes in a cubic crystal.

Figure 6 shows the Miller indices of important planes in a cubic crystal.[§] Some other conventions are the following:

1. $(\bar{h}kl)$: For a plane that intercepts the x -axis on the negative side of the origin, such as $(\bar{1}00)$.
2. $\{hkl\}$: For planes of equivalent symmetry, such as $\{100\}$ for (100) , (010) , (001) , $(\bar{1}00)$, $(0\bar{1}0)$, and $(00\bar{1})$ in cubic symmetry.
3. $[hkl]$: For a crystal direction, such as $[100]$ for the x -axis. By definition, the $[100]$ -direction is perpendicular to (100) -plane, and the $[111]$ -direction is perpendicular to the (111) -plane.
4. $\langle hkl \rangle$: For a full set of equivalent directions, such as $\langle 100 \rangle$ for $[100]$, $[010]$, $[001]$, $[\bar{1}00]$, $[0\bar{1}0]$, and $[00\bar{1}]$.

► 1.3 VALENCE BONDS

As discussed in Section 1.2, each atom in a diamond lattice is surrounded by four nearest neighbors. Figure 7a shows the tetrahedron bonds of a diamond lattice. A simplified two-dimensional bonding diagram for the tetrahedron is shown in Fig. 7b. Each atom has four electrons in the outer orbit, and each atom shares these valence electrons with its four neighbors. This sharing of electrons is known as *covalent bonding*; each *electron pair* constitutes a covalent bond. Covalent bonding occurs between atoms of the same element or between atoms of different elements that have similar outer-shell electron configurations. Each electron spends an equal amount of time with each nucleus. However, both electrons spend most of their time between the two nuclei. The force of attraction for the electrons by both nuclei holds the two atoms together.

Gallium arsenide crystallizes in a zincblende lattice, which also has tetrahedron bonds. The major bonding force in GaAs is also due to the covalent bond. However, gallium arsenide has a small ionic contribution that is an electrostatic attractive force between each Ga^+ ion and its four neighboring As^- ions, or between each As^- ion and its four neighboring Ga^+ ions. Electronically, this means that the paired bonding electrons spend slightly more time in the As atom than in the Ga atom.

At low temperatures, the electrons are bound in their respective tetrahedron lattice; consequently, they are not available for conduction. At higher temperatures, thermal vibrations may break the covalent bonds (ionize one electron from the bond). When a bond is broken, a free electron results and can participate in current conduction.

[§] In Chapter 5, we show that the $\langle 100 \rangle$ orientation is preferred for silicon metal-oxide-semiconductor field-effect transistors (MOSFETs).

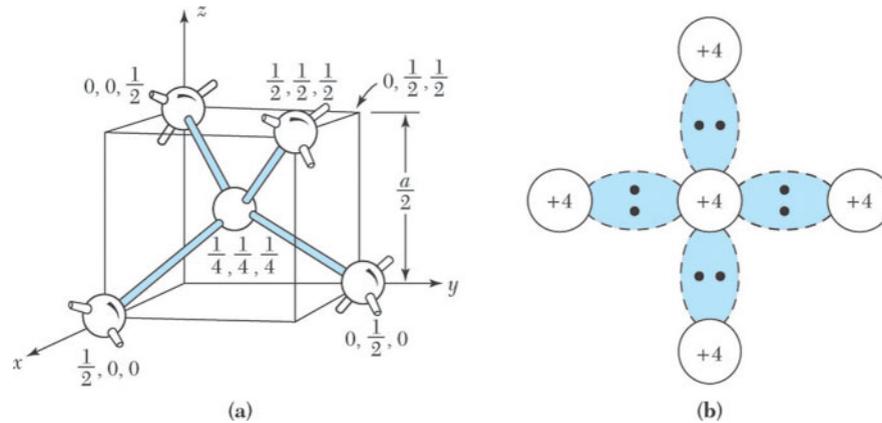


Fig. 7 (a) A tetrahedron bond. (b) Schematic two-dimensional representation of a tetrahedron bond.

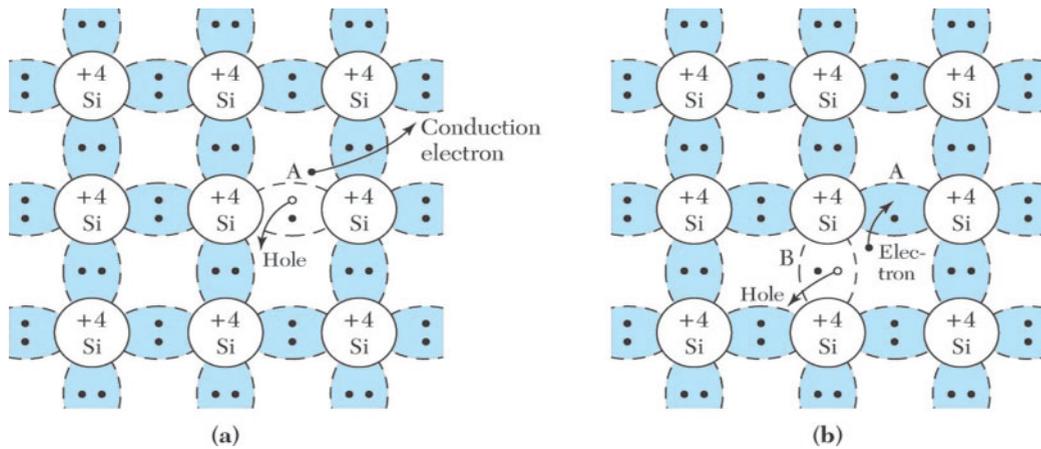


Fig. 8 The basic bond representation of intrinsic silicon. (a) A broken bond at position A, resulting in a conduction electron and a hole. (b) A broken bond at position B.

Figure 8a shows the situation when a valence electron in silicon becomes a free electron. An electron deficiency is left in the covalent bond. This deficiency may be filled by one of the neighboring electrons, which results in a shift of the deficiency location, as from location *A* to location *B* in Fig. 8b. We may, therefore, consider this deficiency as a particle similar to an electron. This fictitious particle is called a *hole*. It carries a positive charge and moves, under the influence of an applied electric field, in the direction opposite to that of an electron. Therefore, both the electron and the hole contribute to the total electric current. The concept of a hole is analogous to that of a bubble in a liquid: although it is actually the liquid that moves, it is much easier to talk about the motion of the bubble in the opposite direction.

► 1.4 ENERGY BANDS

1.4.1 Energy Levels of Isolated Atoms

For an isolated atom, the electrons can have discrete energy levels. For example, the energy levels for an isolated hydrogen atom are given by the Bohr model:⁴

$$E_H = -m_0 q^4 / 8 \epsilon_0^2 h^2 n^2 = -13.6 / n^2 \text{ eV}, \quad (2)$$

where m_0 is the free-electron mass, q is the electronic charge, ϵ_0 is the free-space permittivity, h is the Planck constant, and n is a positive integer called the principal quantum number. The quantity eV (electron volt) is an energy unit corresponding to the energy gained by an electron when its potential is increased by one volt. It is equal to the product of q (1.6×10^{-19} coulomb) and one volt, or 1.6×10^{-19} J. The discrete energies are -13.6 eV for the ground-state energy level ($n = 1$), -3.4 eV for the first excited-state energy level ($n = 2$), and so on. Detailed studies reveal that for higher principle quantum numbers ($n \geq 2$), energy levels are split according to their angular momentum quantum number ($\ell = 0, 1, 2, \dots, n - 1$).

We now consider two identical atoms. When they are far apart, the allowed energy levels for a given principal quantum number (e.g., $n = 1$) consist of one doubly degenerate level; that is, both atoms have exactly the same energy. When they are brought closer, the doubly degenerate energy levels will split into two levels by the interaction between the atoms. The split occurs due to the Pauli exclusion principle, which states that no more than two electrons in a given system can reside in the same energy state at the same time. As N isolated atoms are brought together to form a solid, the orbits of the outer electrons of different atoms overlap and interact with each other. This interaction, including those forces of attraction and repulsion between atoms, causes a shift in the energy levels, as in the case of two interacting atoms. However, instead of two levels, N separate but closely spaced levels are formed. When N is large, the result is an essentially continuous band of energy. This band of N levels can extend over a few eV at the inter-atomic distance of the crystal. The electrons can no longer be treated as belonging to their parent atoms. They belong to the crystal as a whole. Figure 9 shows the effect, where the parameter a represents the equilibrium inter-atomic distance of the crystal.

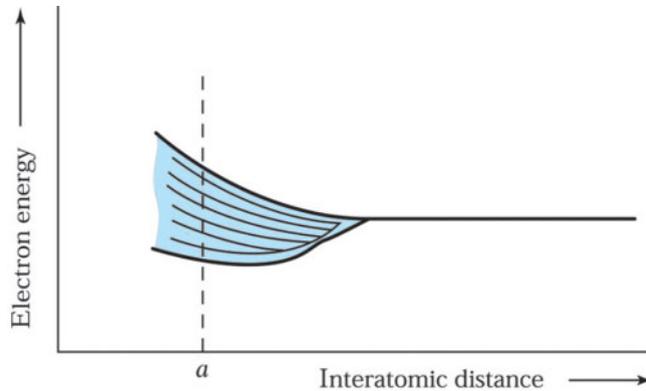


Fig. 9 The splitting of a degenerate state into a band of allowed energies.

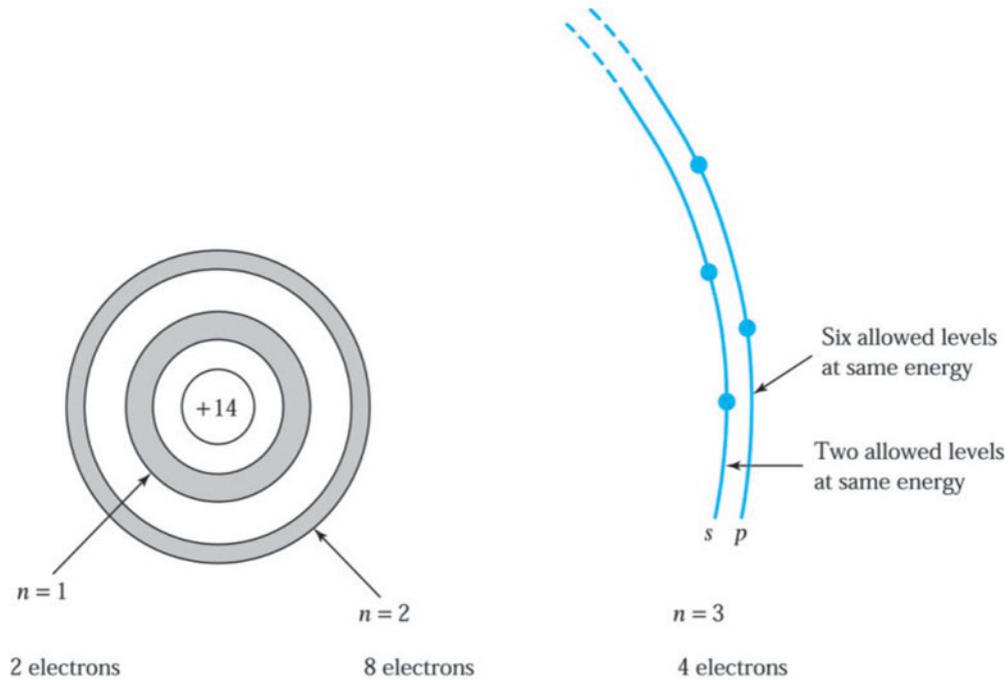


Fig. 10 Schematic representation of an isolated silicon atom.

The actual band splitting in a semiconductor is much more complicated. Figure 10 shows an isolated silicon atom that has 14 electrons. Of the 14 electrons, 10 occupy deep-lying energy levels whose orbital radius is much smaller than the interatomic separation in the crystal. The four remaining valence electrons are relatively weakly bound and can be involved in chemical reactions. Therefore, we only need to consider the outer shell (the $n=3$ level) for the valence electrons, since the two inner shells are completely full and tightly bound to the nucleus. The 3s subshell (i.e., for $n=3$ and $\ell=0$) has two allowed quantum states per atom. This subshell will contain two valence electrons at $T=0$ K. The 3p subshell (i.e., $n=3$ and $\ell=1$) has six allowed quantum states per atom. This subshell will contain the remaining two valence electrons of an individual silicon atom.

Figure 11 is a schematic diagram of the formation of a silicon crystal from N isolated silicon atoms. As the interatomic distance decreases, the 3s and 3p subshell of the N silicon atoms will interact and overlap to form bands. As the 3s and 3p bands grow, they merge into a single band containing $8N$ states. At the equilibrium interatomic distance determined by the condition of minimum total energy, the bands will again split, with $4N$ states in the lower band and $4N$ states in the upper band.

At a temperature of absolute zero, electrons occupy the lowest energy states, so that all states in the lower band (*the valence band*) will be full and all states in the upper band (*the conduction band*) will be empty. The bottom of the conduction band is called E_C and the top of the valence band is called E_V . The *bandgap energy* E_g between the bottom of the conduction band and the top of the valence band ($E_C - E_V$) is the width of the forbidden energy gap, as shown at the far left of Fig. 11. Physically, E_g is the energy required to break a bond in the semiconductor to free an electron to the conduction band and leave a hole in the valence band.

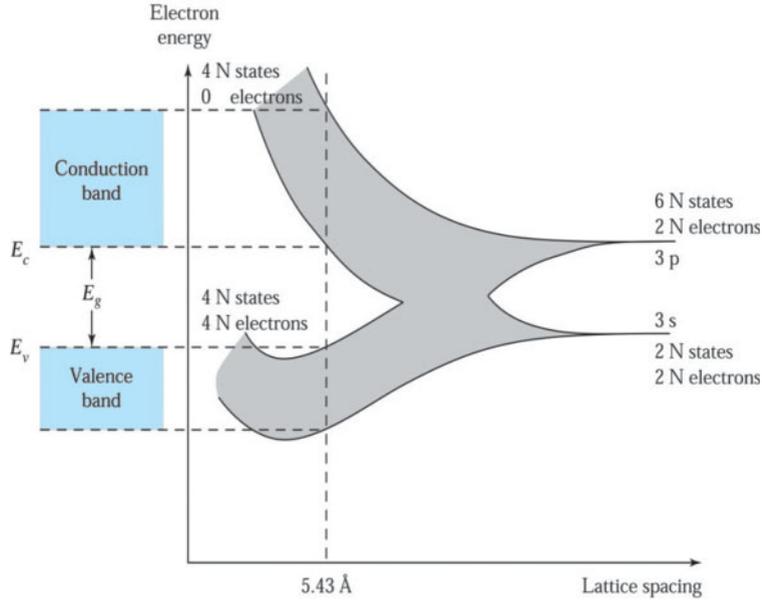


Fig. 11 Formation of energy bands as a diamond lattice crystal is formed by bringing isolated silicon atoms together.

1.4.2 The Energy-Momentum Diagram

The energy E of a free electron is given by

$$E = \frac{p^2}{2m_0}, \quad (3)$$

where p is the momentum and m_0 is the free-electron mass. If we plot E vs. p , we obtain a parabola as shown in Fig. 12. In a semiconductor crystal, an electron in the conduction band is similar to a free electron in being relatively free to move about in the crystal. However, because of the periodic potential of the nuclei, Eq. 3 can no longer be valid. However, it turns out that we can still use Eq. 3 if we replace the free-electron mass in Eq. 3 by an effective mass m_n (the subscript n refers to the negative charge on an electron), that is,

$$E = \frac{p^2}{2m_n}. \quad (4)$$

The electron effective mass depends on the properties of the semiconductor. If we have an energy-momentum relationship described by Eq. 4, we can obtain the effective mass from the second derivative of E with respect to p :

$$m_n \equiv \left(\frac{d^2 E}{dp^2} \right)^{-1}. \quad (5)$$

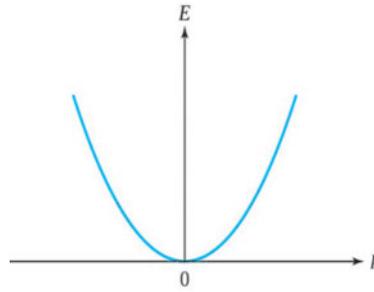


Fig. 12 The parabolic energy (E) vs. momentum (p) curve for a free electron.

Therefore, the narrower the parabola, corresponding to a larger second derivative, the smaller the effective mass. A similar expression can be written for holes (with effective mass m_p where the subscript p refers to the positive charge on a hole). The effective-mass concept is very useful because it enables us to treat electrons and holes essentially as classical charged particles.

Figure 13 shows a simplified energy-momentum relationship of a special semiconductor with an electron effective mass of $m_n = 0.25 m_0$ in the conduction band (the upper parabola) and a hole effective mass of $m_p = m_0$ in the valence band (the lower parabola). Note that the electron energy is measured upward and the hole energy is measured downward. The spacing at $p = 0$ between these two parabolas is the bandgap E_g , shown previously in Fig. 11.

The actual energy-momentum relationships (also called energy-band diagram) for silicon and gallium arsenide are much more complex. Visualized in three dimensions, the relationship between E and p is a complex surface. They are shown in Fig. 14 only for two crystal directions. Since the periodicity of most lattice is different in various directions, the energy-momentum diagram is also different for different directions. In the case of the diamond or zincblende lattice, the maximum in the valence band and minimum in the conduction band occur at $p = 0$ or along one of these two directions. If the minimum of the conduction band occurs at $p = 0$, this means the effective mass of the electrons in every direction in the crystal is the same. It also indicates that the electron motion is independent of crystal direction. If the minimum of the conduction band occurs at $p \neq 0$, this means that the electron behavior in every direction is not the same in the crystal. In general, the minimum of conduction band of polar (with partly ionic binding) semiconductors tend to be at $p = 0$, which is related to the lattice structure and the fraction of ionicity in the bond.

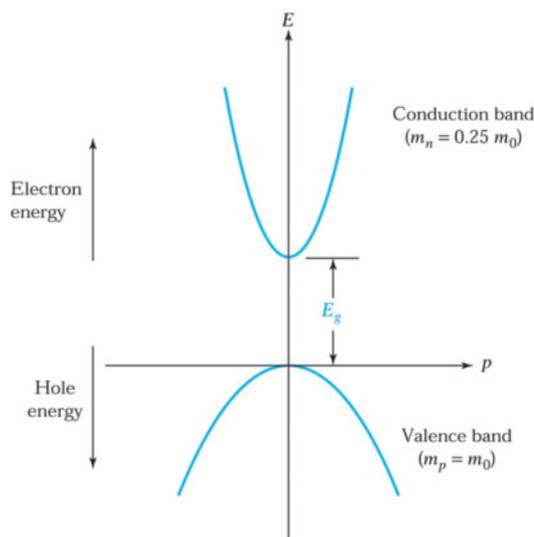


Fig. 13 A schematic energy-momentum diagram for a special semiconductor with $m_n = 0.25 m_0$ and $m_p = m_0$.

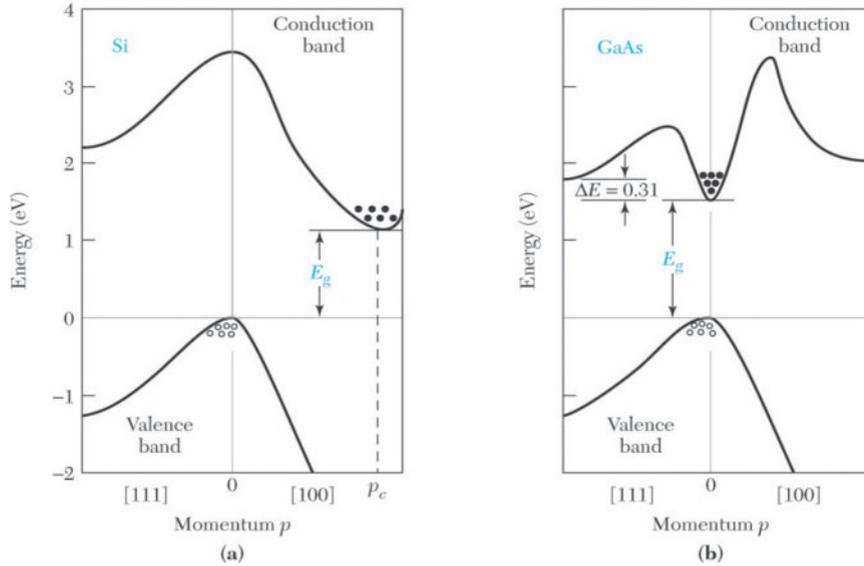


Fig. 14 Energy band structures of (a) Si and (b) GaAs. Circles (o) indicate holes in the valence bands and dots (•) indicate electrons in the conduction bands.

We note that the general features in Fig. 14 are similar to those in Fig. 13. First of all, the valence bands are simpler than the conduction bands. They are qualitatively similar for most semiconductors because the environments for holes moving in the covalent bonds are similar due to the similar structures in diamond and zincblende. There is a bandgap E_g between the bottom of the conduction band and the top of the valence band. Near the minimum of the conduction band or the maximum of the valence band, the E - p curves are essentially parabolic. For silicon (Fig. 14a) the maximum in the valence band occurs at $p = 0$, but the minimum in the conduction band occurs along the [100] direction at $p = p_c$. Therefore, in silicon, when an electron makes a transition from the maximum point in the valence band to the minimum point in the conduction band, not only an energy change ($\geq E_g$) but also some momentum change ($\geq p_c$) is required.

For gallium arsenide (Fig. 14b) the maximum in the valence band and the minimum in the conduction band occur at the same momentum ($p = 0$). Thus, an electron making a transition from the valence band to the conduction band can do so without a change in momentum.

Gallium arsenide is called a *direct semiconductor* because it does not require a change in momentum for an electron transition from the valence band to the conduction band. Silicon is called an *indirect semiconductor* because a change of momentum is required in a transition. This difference between direct and indirect band structures is very important for light-emitting diodes and semiconductor lasers. These devices require direct semiconductors to generate efficiently photons (see Chapters 9 and 10).

We can obtain the effective mass from Fig. 14 using Eq. 5. For example, for gallium arsenide with a very narrow conduction-band parabola, the electron effective mass is $0.063 m_0$, while for silicon, with a wider conduction-band parabola, the electron effective mass is $0.19 m_0$.

1.4.3 Conduction in Metals, Semiconductors, and Insulators

The enormous variation in electrical conductivity of metals, semiconductors, and insulators shown in Fig. 1 may be explained qualitatively in terms of their energy bands. Figure 15 shows the energy band diagrams of three classes of solids—metals, semiconductors, and insulators.

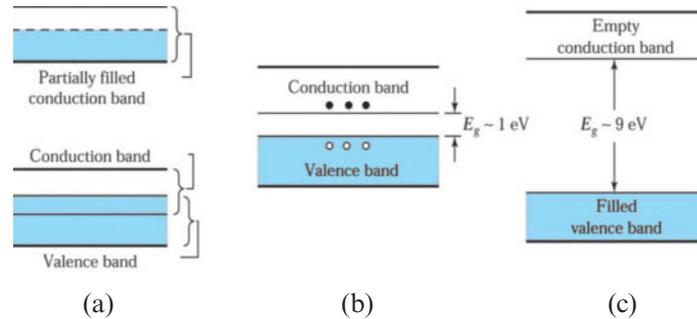


Fig. 15 Schematic energy band representations of (a) a conductor with two possibilities (either the partially filled conduction band shown at the upper portion or the overlapping bands shown at the lower portion), (b) a semiconductor, and (c) an insulator.

Metals

The characteristics of a metal (also called a conductor) include a very low value of resistivity and a conduction band that either is partially filled (as in Cu) or overlaps the valence band (as in Zn or Pb) so that there is no bandgap, as shown in Fig. 15a. As a consequence, the uppermost electrons in the partially filled band or electrons at the top of the valence band can move to the next higher available energy level when they gain kinetic energy (e.g., from an applied electric field). Electrons are free to move with only a small applied field in a metal because there are many unoccupied states close to the occupied energy states. Therefore, current conduction can readily occur in conductors.

Insulators

In an insulator such as silicon dioxide (SiO_2), the valence electrons form strong bonds between neighboring atoms. Since these bonds are difficult to break, there are no free electrons to participate in current conduction at or near room temperature. As shown in the energy band diagram (Fig. 15c), insulators are characterized by a large bandgap. Note that electrons occupy all energy levels in the valence band and all energy levels in the conduction band are empty. Thermal energy[§] or the energy of an applied electric field is insufficient to raise the uppermost electron in the valence band to the conduction band. Thus, although an insulator has many vacant states in the conduction band that can accept electrons, so few electrons actually occupy conduction band states that the overall contribution to electrical conductivity is very small, resulting in a very high resistivity. Therefore, silicon dioxide is an insulator; it can not conduct current.

Semiconductors

Now, consider a material that has a much smaller energy gap, on the order of 1 eV (Fig. 15b). Such materials are called semiconductors. At $T = 0$ K, all electrons are in the valence band, and there are no electrons in the conduction band. Thus, semiconductors are poor conductors at low temperatures. At room temperature and under normal atmospheres, values of E_g are 1.12 eV for Si and 1.42 eV for GaAs. The thermal energy kT at room temperature is a good fraction of E_g , and appreciable numbers of electrons are thermally excited from the valence band to the conduction band. Since there are many empty states in the conduction band, a small applied potential can easily move these electrons, resulting in a moderate current.

► 1.5 INTRINSIC CARRIER CONCENTRATION

We now derive the carrier concentration in the thermal equilibrium condition, that is, the steady-state condition at a given temperature without any external excitations such as light, pressure, or an electric field. At a given temperature, continuous thermal agitation results in the excitation of electrons from the valence band to the conduction band and leaves an equal number of holes in the valence band. An *intrinsic semiconductor* is one that contains relatively small amounts of impurities compared with the thermally generated electrons and holes.

[§]The thermal energy is of the order of kT . At room temperature, kT is 0.026 eV, which is much smaller than the bandgap of an insulator.

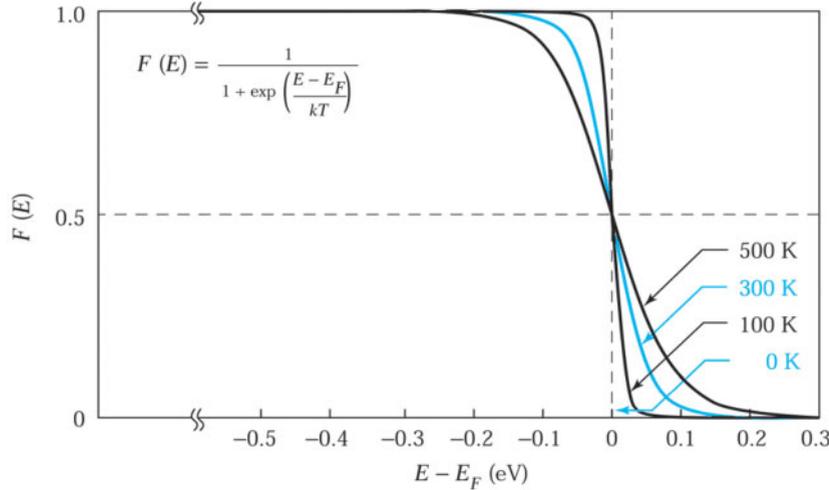


Fig. 16 Fermi distribution function $F(E)$ versus $(E - E_F)$ for various temperatures.

To obtain the electron density (i.e., the number of electrons per unit volume) in an intrinsic semiconductor, we first evaluate the electron density in an incremental energy range dE . This density $n(E)$ is given by the product of the density of states $N(E)$, that is, the density of allowed energy states (including electron spin) per energy range per unit volume,[§] and by the probability of occupying that energy range $F(E)$. Thus, the electron density in the conduction band is given by integrating $N(E) F(E) dE$ from the bottom of the conduction band (E_c initially taken to be $E = 0$ for simplicity) to the top of the conduction band E_{top} :

$$n = \int_0^{E_{top}} n(E) dE = \int_0^{E_{top}} N(E) F(E) dE, \tag{6}$$

where n is in cm^{-3} , and $N(E)$ is in $(\text{cm}^3\text{-eV})^{-1}$.

The probability that an electron occupies an electronic state with energy E is given by the Fermi–Dirac distribution function, which is also called the Fermi distribution function

$$F(E) = \frac{1}{1 + e^{(E - E_F)/kT}}, \tag{7}$$

where k is the Boltzmann constant, T is the absolute temperature in degrees Kelvin, and E_F is the energy of the Fermi level. The Fermi level is the energy at which the probability of occupation by an electron is exactly one-half. The Fermi distribution is illustrated in Fig. 16 for different temperatures. Note that $F(E)$ is symmetrical around the Fermi level E_F .

For energies that are $3kT$ above or below the Fermi energy, the exponential term in Eq. 7 becomes larger than 20 or smaller than 0.05, respectively. The Fermi distribution function can thus be approximated by simpler expressions:

[§] The density of states $N(E)$ is derived in Appendix H.

$$F(E) \cong e^{-(E - E_F)/kT} \quad \text{for } (E - E_F) > 3kT, \quad (8a)$$

and

$$F(E) \cong 1 - e^{-(E - E_F)/kT} \quad \text{for } (E - E_F) < 3kT \quad (8b)$$

Equation 8b can be regarded as the probability that a hole occupies a state located at energy E .

Figure 17 shows schematically from left to right the band diagram, the density of states $N(E)$, which varies as \sqrt{E} for a given electron effective mass, the Fermi distribution function, and the carrier concentrations for an intrinsic semiconductor. The electron concentration can be obtained graphically from Fig. 17 using Eq. 6; that is, the product of $N(E)$ in Fig. 17b and $F(E)$ in Fig. 17c gives the $n(E)$ -versus- E curve (upper curve) in Fig. 17d. The upper shaded area in Fig. 17d corresponds to the electron density.

There are a large number of allowed states in the conduction band. However, for an intrinsic semiconductor there will not be many electrons in the conduction band. Therefore, the probability of an electron occupying one of these states is small. Also, there are a large number of allowed states in the valence band. By contrast, most of these are occupied by electrons. Thus, the probability of an electron occupying one of these states in the valence band is nearly unity. There will be only a few unoccupied electron states, that is, holes, in the valence band. From Fig. 16 then, all electrons are in the valence band, and there are no electrons in the conduction band at $T = 0$ K. The Fermi energy E_F for which the probability of occupation by an electron is 0.5 lies midway between the two bands. At a finite temperature, the number of electrons in the conduction band is equal to the number of holes in the valence band. The Fermi distribution is symmetrical around the Fermi level E_F . The Fermi level must be at the midgap in order to obtain equal electron and hole concentrations if the density of state in the conduction and valence bands is the same. That is to say, E_F is independent of temperature for an intrinsic semiconductor. As can be seen, the Fermi level is located near the middle of the bandgap. Substituting the last equation in Appendix H and Eq. 8a into Eq. 6 yields[§]

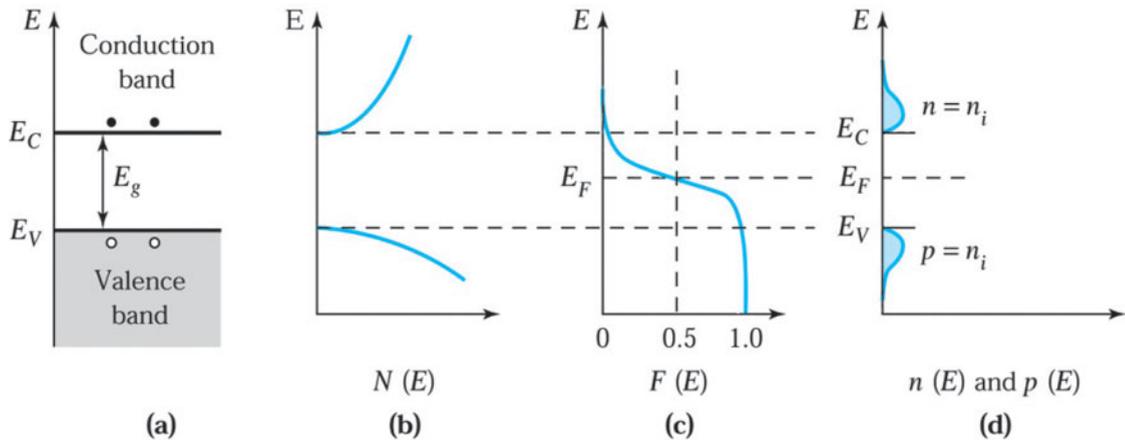


Fig. 17 Intrinsic semiconductor. (a) Schematic band diagram. (b) Density of states. (c) Fermi distribution function. (d) Carrier concentration.

[§] We have taken E_{top} to be ∞ , because $F(E)$ becomes very small when $(E - E_C) \gg kT$.

$$n = \frac{2}{\sqrt{\pi}} N_C (kT)^{-3/2} \int_0^{\infty} E^{1/2} \exp[-(E - E_F)/kT] dE, \quad (9)$$

where $N_C \equiv 12(2\pi m_n kT/h^2)^{3/2}$ for Si (10a)

$$\equiv 2(2\pi m_n kT/h^2)^{3/2} \text{ for GaAs.} \quad (10b)$$

If we let $x \equiv E/kT$, Eq. 9 becomes

$$n = \frac{2}{\sqrt{\pi}} N_C \exp(E_F/kT) \int_0^{\infty} x^{1/2} e^{-x} dx. \quad (11)$$

The integral in Eq. 11 is of the standard form and equals $\sqrt{\pi}/2$. Therefore, Eq. 11 becomes

$$n = N_C \exp(E_F/kT). \quad (12)$$

If we refer to the bottom of the conduction band as E_c instead of $E = 0$, we obtain for the electron density in the conduction band

$$n = N_C \exp[(E_c - E_F)/kT], \quad (13)$$

where N_C defined in Eq.10 is the *effective density of states* in the conduction band. At room temperature (300 K), N_C is $2.86 \times 10^{19} \text{ cm}^{-3}$ for silicon and $4.7 \times 10^{17} \text{ cm}^{-3}$ for gallium arsenide.

Similarly, we can obtain the hole density p in the valence band:

$$p = N_V \exp[(E_F - E_V)/kT], \quad (14)$$

and $N_V \equiv 2(2\pi m_p kT/h^2)^{3/2}, \quad (15)$

where N_V is the *effective density of states in the valence band* for both Si and GaAs. At room temperature, N_V is $2.66 \times 10^{19} \text{ cm}^{-3}$ for silicon and $7.0 \times 10^{18} \text{ cm}^{-3}$ for gallium arsenide.

For an intrinsic semiconductor, the number of electrons per unit volume in the conduction band is equal to the number of holes per unit volume in the valence band, that is, $n = p = n_i$ where n_i is the *intrinsic carrier density*. This relationship of electrons and holes is depicted in Fig. 17d. Note that the shaded area in the conduction band is the same as that in the valence band.

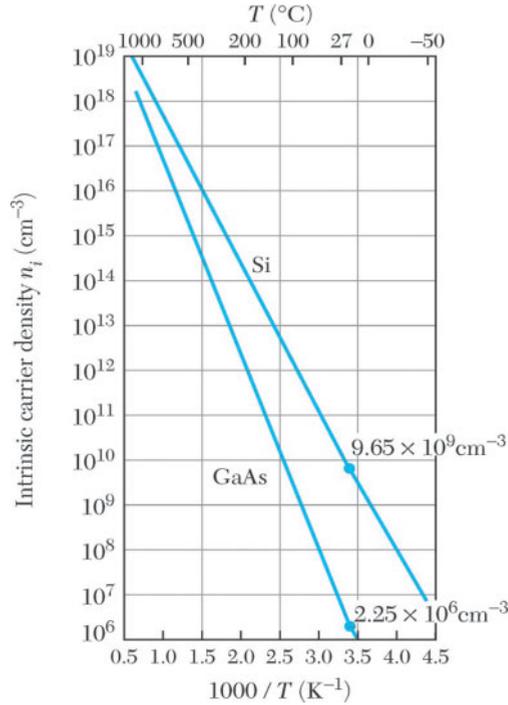


Fig. 18 Intrinsic carrier densities in Si and GaAs as a function of the reciprocal of temperature.⁵⁻⁷

The Fermi level for an intrinsic semiconductor is obtained by equating Eq.13 and Eq. 14:

$$E_F = E_i = (E_C + E_V)/2 + (kT/2) \ln (N_V/N_C). \quad (16)$$

At room temperature, the second term is much smaller than the bandgap. Hence, the intrinsic Fermi level E_i of an intrinsic semiconductor generally lies very close to the middle of the bandgap.

The intrinsic carrier density is obtained from Eqs. 13, 14, and 16:

$$np = n_i^2, \quad (17)$$

$$n_i^2 = N_C N_V \exp(-E_g / kT), \quad (18)$$

and

$$n_i = \sqrt{N_C N_V} \exp(-E_g / 2kT), \quad (19)$$

where $E_g \equiv E_C - E_V$. Figure 18 shows the temperature dependence of n_i for silicon and gallium arsenide.⁵ At room temperature (300 K), n_i is $9.65 \times 10^9 \text{ cm}^{-3}$ for silicon⁶ and $2.25 \times 10^6 \text{ cm}^{-3}$ for gallium arsenide.⁷ As expected, the larger the bandgap, the smaller the intrinsic carrier density.

► 1.6 DONORS AND ACCEPTORS

When a semiconductor is doped with impurities, the semiconductor becomes *extrinsic* and impurity energy levels are introduced. Figure 19a shows schematically that a silicon atom is replaced (or substituted) by an arsenic atom with five valence electrons. The arsenic atom forms covalent bonds with its four neighboring silicon atoms. The fifth electron has a relatively small binding energy to its host arsenic atom and can be “ionized” to become a conduction electron at a moderate temperature. We say that this electron has been “donated” to the conduction band. The arsenic atom is called a *donor* and the silicon becomes *n*-type because of the addition of the negative charge carrier. Similarly, Fig. 19b shows that when a boron atom with three valence electrons substitutes for a silicon atom, an additional electron is “accepted” to form four covalent bonds around the boron, and a positively charged “hole” is created in the valence band. This is a *p*-type semiconductor, and the boron is an *acceptor*.

The impurity atoms are imperfections and interrupt the perfect periodicity of the lattice; energy levels within the band gap that were forbidden are no longer disallowed. That is to say, the impurity atoms will introduce an energy level or multiple energy levels in the band gap.

We can estimate the *ionization energy* for the donor E_D by replacing m_0 with the electron effective mass m_n and taking into account the semiconductor permittivity ϵ_s in the hydrogen atom model, Eq. 2:

$$E_D = \left(\frac{\epsilon_0}{\epsilon_s} \right)^2 \left(\frac{m_n}{m_0} \right) E_H. \quad (20)$$

The ionization energy for donors, measured from the conduction band edge and calculated from Eq. 20 is 0.025 eV for silicon and 0.007 eV for gallium arsenide. The hydrogen atom calculation for the ionization level of acceptors is similar to that for donors. We consider the unfilled valence band as a filled band plus a hole in the central force field of a negative charged acceptor. The calculated ionization energy, measured from the valence band edge, is 0.05 eV for both silicon and gallium arsenide.

This simple hydrogen atom model cannot account for the details of the ionization energy, particularly for the deep impurity levels in semiconductors (i.e., with ionization energies $\geq 3 kT$). However, the calculated values do predict the correct order of magnitude of the true ionization energies for shallow impurity levels. Figure 20 shows the measured ionization energies for various impurities in silicon and gallium arsenide.⁸ Note that it is possible for a single atom to have many levels; for example, oxygen in silicon has two donor levels and two acceptor levels in the forbidden energy gap.

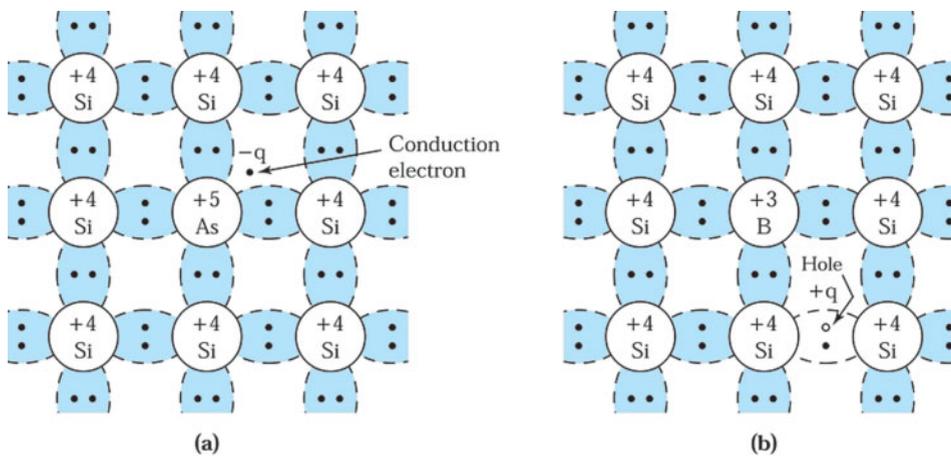


Fig. 19 Schematic bond pictures for (a) *n*-type Si with donor (arsenic) and (b) *p*-type Si with acceptor (boron).

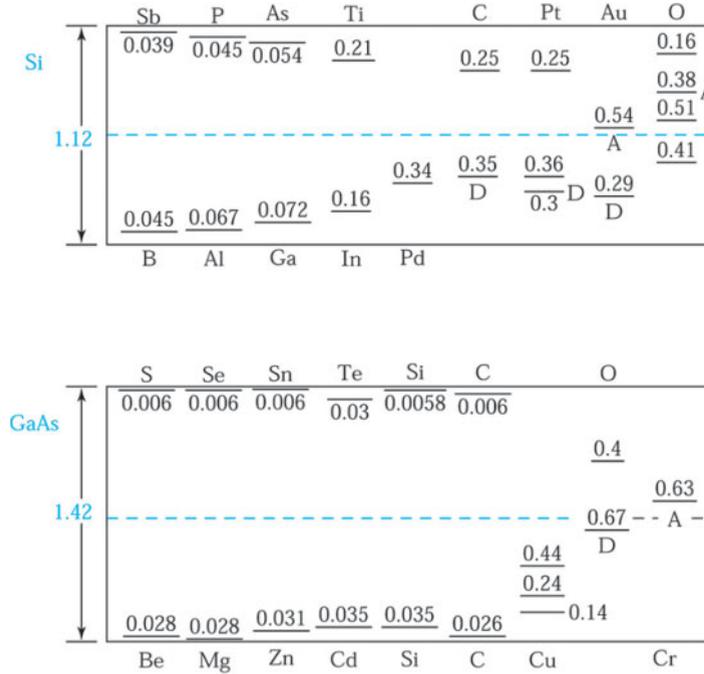


Fig. 20 Measured ionization energies (in eV) for various impurities in Si and GaAs. The levels below the gap center are measured from the top of the valence band and are acceptor levels unless labeled by *D* for donor level. The levels above the gap center are measured from the bottom of the conduction band and are donor levels unless indicated by *A* for acceptor level.⁸

1.6.1 Nondegenerate Semiconductor

In our previous discussion, we have assumed that the electron or hole concentration is much lower than the effective density of states in the conduction band or the valence band, respectively. In other words, the Fermi level E_F is at least $3kT$ above E_V or $3kT$ below E_C . In such cases, the semiconductor is referred to as a *nondegenerate* semiconductor.

For shallow donors in *n*-type silicon and gallium arsenide, there usually is enough thermal energy to supply the energy E_D to ionize all donor impurities at room temperature and thus provide the same number of electrons in the conduction band. This condition is called complete ionization. Under a complete ionization condition, we can write the electron density as

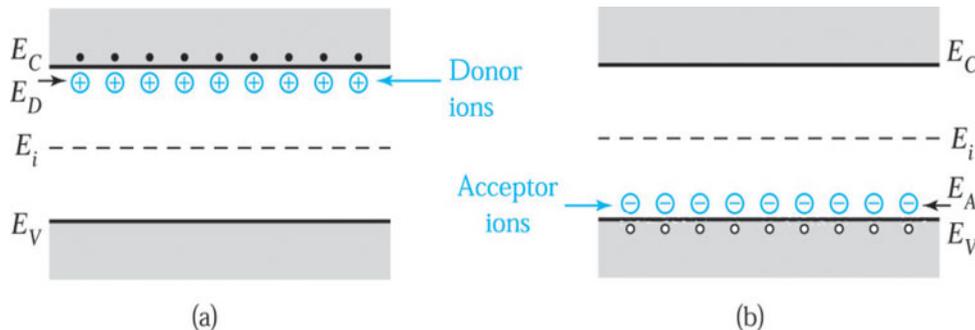


Fig. 21 Schematic energy band representation of extrinsic semiconductors with (a) donor ions and (b) acceptor ions.

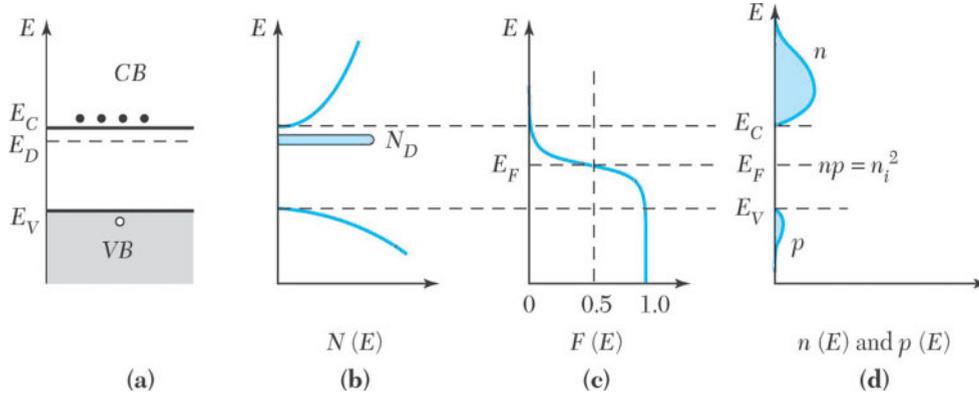


Fig. 22 *n*-Type semiconductor. (a) Schematic band diagram. (b) Density of states. (c) Fermi distribution function. (d) Carrier concentration. Note that $np = n_i^2$.

$$n = N_D \quad (21)$$

where N_D is the donor concentration. Figure 21a illustrates complete ionization where the donor level E_D is measured with respect to the bottom of the conduction band and equal concentrations of electrons (which are mobile) and donor ions (which are immobile) are shown. From Eqs. 13 and 21, we obtain the Fermi level in terms of the effective density of states N_C and the donor concentration N_D :

$$E_C - E_F = kT \ln(N_C/N_D). \quad (22)$$

Similarly, for shallow acceptors for *p*-type semiconductors as shown in Fig. 21b, if there is complete ionization, the concentration of holes is

$$p = N_A, \quad (23)$$

where N_A is the acceptor concentration. We can obtain the corresponding Fermi level from Eqs. 14 and 23:

$$E_F - E_V = kT \ln(N_V/N_A). \quad (24)$$

From Eq. 22 we can see that the higher the donor concentration, the smaller the energy difference ($E_C - E_F$); that is, the Fermi level will move closer to the bottom of the conduction band. Similarly, for higher acceptor concentration, the Fermi level will move closer to the top of the valence band. Figure 22 illustrates the procedure for obtaining the carrier concentrations for an *n*-type semiconductor. This figure is similar to that shown in Fig. 17. However, the Fermi level is closer to the bottom of the conduction band, and the electron concentration (upper shaded area) is much larger than the hole concentration (lower shaded area).

It is useful to express electron and hole densities in terms of the intrinsic carrier concentration n_i and the intrinsic Fermi level E_i since E_i is frequently used as a reference level when discussing extrinsic semiconductors. From Eq.13 we obtain

$$n = N_C \exp\left[-(E_C - E_F)/kT \right],$$

$$N_C \exp\left[-(E_C - E_i)/kT \right] \exp\left[(E_F - E_i)/kT \right],$$

or $n = n_i \exp\left[(E_F - E_i)/kT \right],$ (25)

and similarly,

$$p = n_i \exp[(E_i - E_F) / kT]. \quad (26)$$

Note that the product of n and p from Eqs. 25 and 26 is n_i^2 . This result is identical to that for the intrinsic case, Eq. 17. Equation 17 is called the *mass action law*, and is valid for both intrinsic and extrinsic semiconductors under thermal equilibrium conduction. In an extrinsic semiconductor, the Fermi level moves toward either the bottom of the conduction band (n -type) or the top of the valence band (p -type). Either n - or p -type carriers will then dominate, but the product of the two types of carriers will remain constant at a given temperature.

► EXAMPLE 4

A silicon ingot is doped with 10^{16} arsenic atoms/cm³. Find the carrier concentrations and the Fermi level at room temperature (300 K).

SOLUTION At 300 K, we can assume complete ionization of impurity atoms. We have

$$n \approx N_D = 10^{16} \text{ cm}^{-3}.$$

From Eq. 17,

$$p \approx n_i^2 / N_D = (9.65 \times 10^9)^2 / 10^{16} = 9.3 \times 10^3 \text{ cm}^{-3}.$$

The Fermi level measured from the bottom of the conduction band is given by Eq. 22:

$$\begin{aligned} E_C - E_F &= kT \ln(N_C / N_D) \\ &= 0.0259 \ln(2.86 \times 10^{19} / 10^{16}) = 0.205 \text{ eV}. \end{aligned}$$

The Fermi level measured from the intrinsic Fermi level is given by Eq. 25:

$$\begin{aligned} E_F - E_i &\approx kT \ln(N_D / n_i) \\ &= 0.0259 \ln(10^{16} / 9.65 \times 10^9) = 0.358 \text{ eV}. \end{aligned}$$

These results are shown graphically in Fig. 23. ◀

If both donor and acceptor impurities are present simultaneously, the impurity that is present in a greater concentration determines the type of conductivity in the semiconductor.

The Fermi level must adjust itself to preserve charge neutrality, that is, the total negative charges (electrons and ionized acceptors) must equal the total positive charges (holes and ionized donors). Under complete ionization condition, we have

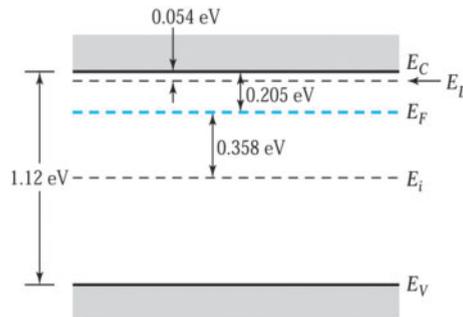


Fig. 23 Band diagram showing Fermi level E_F and intrinsic Fermi level E_i .

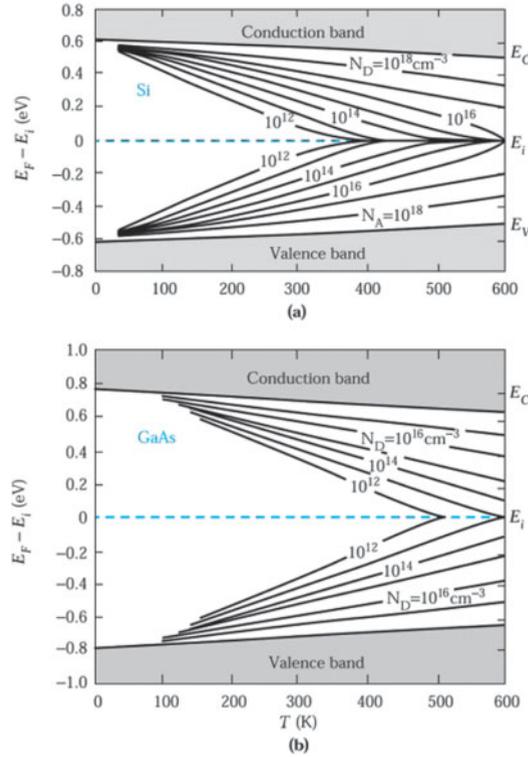


Fig. 24 Fermi level for (a) Si and (b) GaAs as a function of temperature and impurity concentration showing dependence of the bandgap on temperature.⁹

$$\boxed{n + N_A = p + N_D.} \quad (27)$$

Solving Eqs. 17 and 27 yields the equilibrium electron and hole concentrations in an n -type semiconductor:

$$\boxed{n_n = \frac{1}{2} \left[N_D - N_A + \sqrt{(N_D - N_A)^2 + 4n_i^2} \right],} \quad (28)$$

$$\boxed{p_n = n_i^2 / n_n.} \quad (29)$$

The subscript n refers to the n -type semiconductor. Because the electron is the dominant carrier, it is called the *majority carrier*. The hole in the n -type semiconductor is called the *minority carrier*. Similarly, we obtain the concentration of holes (majority carrier) and electrons (minority carrier) in a p -type semiconductor as:

$$\boxed{p_p = \frac{1}{2} \left[N_A - N_D + \sqrt{(N_D - N_A)^2 + 4n_i^2} \right],} \quad (30)$$

$$\boxed{n_p = n_i^2 / p_p.} \quad (31)$$

The subscript p refers to the p -type semiconductor.

Generally, the magnitude of the net impurity concentration $|N_D - N_A|$ is greater than the intrinsic carrier concentration n_i ; therefore, the above relationships can be simplified to

$$n_n \approx N_D - N_A \quad \text{if} \quad N_D > N_A, \quad (32)$$

$$p_p \approx N_A - N_D \quad \text{if} \quad N_A > N_D. \quad (33)$$

From Eqs. 28 to 31 together with Eqs. 13 and 14, we can calculate the position of the Fermi level as a function of temperature for a given acceptor or donor concentration. Figure 24 plots these calculations for silicon⁹ and gallium arsenide. We have incorporated in the figure the variation of the bandgap with temperature (see Problem 7). Note that as the temperature increases, the Fermi level approaches the intrinsic level, that is, the semiconductor becomes intrinsic.

Figure 25 shows electron density in Si as a function of temperature for a donor concentration of $N_D = 10^{15} \text{ cm}^{-3}$. At low temperatures, the thermal energy in the crystal is not sufficient to ionize all the donor impurities present. Some electrons are “frozen” at the donor level and the electron density is less than the donor concentration. As the temperature is increased, the condition of complete ionization is reached, (i.e., $n_n = N_D$). As the temperature is further increased, the electron concentration remains essentially the same over a wide temperature range. This is the extrinsic region. However, as the temperature is increased even further, we reach a point where the intrinsic carrier concentration becomes comparable to the donor concentration. Beyond this point, the semiconductor becomes intrinsic. The temperature at which the semiconductor becomes intrinsic depends on the impurity concentrations and the bandgap value and can be obtained from Fig. 18 by setting the impurity concentration equal to n_i .

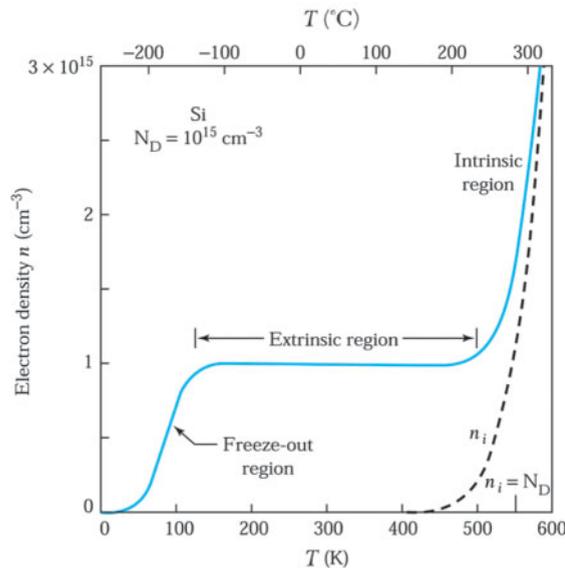


Fig. 25 Electron density as a function of temperature for a Si sample with a donor concentration of 10^{15} cm^{-3} .

1.6.2 Degenerate Semiconductor

When the doping concentration becomes equal or larger than the corresponding effective density of states, we can no longer use the approximation of Eq. 8, and the electron density (Eq. 6) has to be integrated numerically. For very heavily doped n -type or p -type semiconductor, E_F will be above E_C or below E_V . The semiconductor is referred to as a *degenerate* semiconductor.

An important aspect of high doping is the bandgap-narrowing effect; that is, high impurity concentration causes a reduction of the bandgap. The bandgap reduction ΔE_g for silicon at room temperature is given by

$$\Delta E_g \approx 22 \left(\frac{N}{10^{18}} \right)^{1/2} \text{ meV}, \quad (34)$$

where the doping is in cm^{-3} . For example, for $N_D \leq 10^{18} \text{ cm}^{-3}$, $\Delta E_g \leq 0.022 \text{ eV}$, which is less than 2% of the original bandgap. However, for $N_D \geq N_C = 2.86 \times 10^{19} \text{ cm}^{-3}$, $\Delta E_g \geq 0.12 \text{ eV}$, which is a significant fraction of E_g .

► SUMMARY

At the beginning of the chapter we listed a few important semiconductor materials. The properties of semiconductors are determined to a large extent by the crystal structure. We have defined the Miller indices to describe the crystal surfaces and crystal orientations. A discussion of how to grow semiconductor crystals can be found in Chapter 11.

The bonding of atoms and the electron energy-momentum relationship in a semiconductor were considered in connection with the electrical properties. The energy band diagram can be used to understand why some materials are good conductors of electric current whereas others are poor conductors. We have also shown that changing the temperature or the amount of impurities can drastically vary the conductivity of a semiconductor.

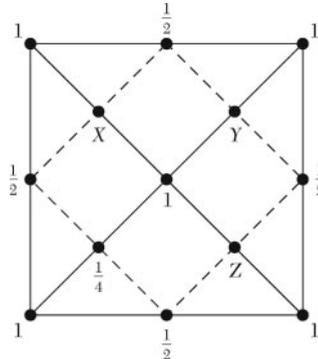
► REFERENCES

1. R. A. Smith, *Semiconductors*, 2nd ed., Cambridge Univ. Press, London, 1979.
2. R. F. Pierret, *Semiconductor Device Fundamentals*, Addison Wesley, Boston, MA, 1996.
3. C. Kittel, *Introduction to Solid State Physics*, 6th ed., Wiley, New York, 1986.
4. D. Halliday and R. Resnick, *Fundamentals of Physics*, 2nd ed., Wiley, New York, 1981.
5. C. D. Thurmond, "The Standard Thermodynamic Function of the Formation of Electrons and Holes in Ge, Si, GaAs, and GaP," *J. Electrochem. Soc.*, **122**, 1133 (1975).
6. P. P. Altermatt, et al., "The Influence of a New Bandgap Narrowing Model on Measurement of the Intrinsic Carrier Density in Crystalline Silicon," *Tech. Dig., 11th Int. Photovoltaic Sci. Eng. Conf.*, Sapporo, p. 719 (1999).
7. J. S. Blackmore, "Semiconducting and Other Major Properties of Gallium Arsenide," *J. Appl. Phys.*, **53**, 123–181 (1982).
8. S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, 3rd ed., Wiley Interscience, Hoboken, 2007.
9. A. S. Grove, *Physics and Technology of Semiconductor Devices*, Wiley, New York, 1967.

► PROBLEMS (* DENOTES DIFFICULT PROBLEMS)

FOR SECTION 1.2 BASIC CRYSTAL STRUCTURES

- (a) What is the distance between nearest neighbors in silicon?
(b) Find the number of atoms per square centimeter in silicon in the (100), (110), and (111) planes.
- If we project the atoms in a diamond lattice onto the bottom surface with the heights of the atoms in unit of the lattice constant shown in the figure below; find the heights of the three atoms (X, Y, Z) on the figure.



- Find the maximum fraction of the unit cell volume, which can be filled by identical hard spheres in the simple cubic, face-centered cubic, and diamond lattices.
- Calculate the tetrahedral bond angle, the angle between any pair of the four bonds in a diamond lattice. (Hint: represent the four bonds as vectors of equal length. What must the sum of the four vectors equal? Take components of this vector equation along the direction of one of these vectors.)
- If a plane has intercepts at $2(a)$, $3(a)$, and $4(a)$ along the three Cartesian coordinates, where a is the lattice constant, find the Miller indices of the planes.
- (a) Calculate the density of GaAs (the lattice constant of GaAs is 5.65 \AA , and the atomic weights of Ga and As are 69.72 and 74.92 g/mol , respectively).
(b) A gallium arsenide sample is doped with tin. If the tin displaces gallium atoms in crystal lattice, are donors or acceptors formed? Why? Is the semiconductor n - or p -type?

FOR SECTION 1.4 ENERGY BANDS

- The variation of silicon and GaAs bandgaps with temperature can be expressed as $E_g(T) = E_g(0) - \alpha T^2/(T + \beta)$, where $E_g(0) = 1.17 \text{ eV}$, $\alpha = 4.73 \times 10^{-4} \text{ eV/K}$, and $\beta = 636 \text{ K}$ for silicon; and $E_g(0) = 1.519 \text{ eV}$, $\alpha = 5.405 \times 10^{-4} \text{ eV/K}$, and $\beta = 204 \text{ K}$ for GaAs. Find the bandgaps of Si and GaAs at 100 K and 600 K .

FOR SECTION 1.5 INTRINSIC CARRIER CONCENTRATION

- Derive Eq. 14. (Hint: In the valence band, the probability of occupancy of a state by a hole is $[1 - F(E)]$.)
- At room temperature (300 K) the effective density of states in the valence band is $2.66 \times 10^{19} \text{ cm}^{-3}$ for silicon and $7 \times 10^{18} \text{ cm}^{-3}$ for gallium arsenide. Find the corresponding effective masses of holes. Compare these masses with the free-electron mass.

42 Semiconductors

10. Calculate the location of E_i in silicon at liquid nitrogen temperature (77 K), at room temperature (300 K), and at 100°C (let $m_p = 1.0 m_0$ and $m_n = 0.19 m_0$). Is it reasonable to assume that E_i is in the center of the forbidden gap?
11. Find the kinetic energy of electrons in the conduction band of a nondegenerate n -type semiconductor at 300 K.
12. (a) For a free electron with a velocity of 10^7 cm/s, what is its de Broglie wavelength? (b) In GaAs, the effective mass of electrons in the conduction band is $0.063 m_0$. If they have the same velocity, find the corresponding de Broglie wavelength.
13. The intrinsic temperature of a semiconductor is the temperatures at which the intrinsic carrier concentration equals the impurity concentration. Find the intrinsic temperature for a silicon sample doped with 10^{15} phosphorus atoms/cm³.

FOR SECTION 1.6 DONORS AND ACCEPTORS

14. A silicon sample at $T = 300$ K contains an acceptor impurity concentration of $N_A = 10^{16}$ cm⁻³. Determine the concentration of donor impurity atoms that must be added so that the silicon is n -type and the Fermi energy is 0.20 eV below the conduction band edge.
15. Draw a simple flat energy band diagram for silicon doped with 10^{16} arsenic atoms/cm³ at 77 K, 300 K, and 600 K. Show the Fermi level and use the intrinsic Fermi level as the energy reference.
16. Find the electron and hole concentrations and Fermi level in silicon at 300 K (a) for 1×10^{15} boron atoms/cm³ and (b) for 3×10^{16} boron atoms/cm³ and 2.9×10^{16} arsenic atoms/cm³.
17. A Si sample is doped with 10^{17} As atoms/cm³. What is the equilibrium hole concentration p_0 at 300 K? Where is E_F relative to E_i ?
18. A Ge sample is doped with 2.5×10^{13} cm⁻³ donor atoms that can be assumed to be fully ionized at room temperature. What is the free electron concentration for this sample at room temperature? (The intrinsic carrier concentration n_i of Ge is 2.5×10^{13} cm⁻³.)
19. A p -type Si is doped with N_A acceptors close to the valence band edge. A certain type of donor impurity whose energy level is located at the intrinsic level is to be added to the semiconductor to obtain perfect compensation. If we assume that simple Fermi-level statistics apply, what is the concentration of donors required? Furthermore, after adding the donor impurity, what is the total number of ionized impurities if the above sample is perfect compensation?
20. Calculate the Fermi level of silicon doped with 10^{15} , 10^{17} , and 10^{19} phosphorus atoms/cm³ at room temperature, assuming complete ionization. From the calculated Fermi level, check if the assumption of complete ionization is justified for each doping. Assume that the ionized donors is given by $n = N_D [1 - F(E_D)] = \frac{N_D}{1 + \exp[(E_F - E_D) / kT]}$.
21. For an n -type silicon sample with 10^{16} cm⁻³ phosphorous donor impurities and a donor level at $E_D = 0.045$ eV, find the ratio of the neutral donor density to the ionized donor density at 77 K where the Fermi level is 0.0459 below the bottom of the conduction band. The expression for ionized donors is given in Prob. 20.

Carrier Transport Phenomena

- ▶ 2.1 CARRIER DRIFT
 - ▶ 2.2 CARRIER DIFFUSION
 - ▶ 2.3 GENERATION AND RECOMBINATION PROCESSES
 - ▶ 2.4 CONTINUITY EQUATION
 - ▶ 2.5 THERMIONIC EMISSION PROCESS
 - ▶ 2.6 TUNNELING PROCESS
 - ▶ 2.7 SPACE CHARGE EFFECT
 - ▶ 2.8 HIGH-FIELD EFFECTS
 - ▶ SUMMARY
-

In this chapter, we investigate various transport phenomena in semiconductor devices. The transport processes include drift, diffusion, recombination, generation, thermionic emission, space charge effect, tunneling, and impact ionization. We consider the motion of charge carriers (electrons and holes) in semiconductors under the influence of both an electric field and a carrier concentration gradient. We also discuss the concept of the nonequilibrium condition, where the carrier concentration product pn is different from its equilibrium value n_i^2 . Returning to an equilibrium condition through the generation-recombination processes will be discussed next. We then derive the basic governing equations for semiconductor device operation, which includes the current density equation and the continuity equation. This is followed by a discussion of thermionic emission, the tunneling process, and the space-charge effect. The chapter closes with a brief discussion of high-field effects, including velocity saturation and impact ionization.

Specifically, we cover the following topics:

- The current density equation and its drift and diffusion components.
- The continuity equation and its generation and recombination components.
- Other transport phenomena, including thermionic emission, tunneling, space-charge effect, transferred-electron effect, and impact ionization.
- Methods to measure key semiconductor parameters such as resistivity, mobility, majority-carrier concentration, and minority-carrier lifetime.

▶ 2.1 CARRIER DRIFT

2.1.1 Mobility

Consider an n -type semiconductor sample with uniform donor concentration in thermal equilibrium. As discussed in Chapter 1, the conduction electrons in the semiconductor conduction band are essentially free particles, since they are not associated with any particular lattice or donor site. The influence of crystal lattices is incorporated

in the effective mass of conduction electrons, which differs somewhat from the mass of free electrons. Under thermal equilibrium, the average thermal energy of a conduction electron can be obtained from the theorem for equipartition of energy, $1/2 kT$ units of energy per degree of freedom, where k is Boltzmann's constant and T is the absolute temperature. The electrons in a semiconductor have three degrees of freedom; they can move around in a three-dimensional space. Therefore, the kinetic energy of the electrons is given by

$$\frac{1}{2} m_n v_{th}^2 = \frac{3}{2} kT, \quad (1)$$

where m_n is the effective mass of electrons and v_{th} is the average thermal velocity. At room temperature (300 K), the thermal velocity of electrons in Eq. 1 is about 10^7 cm/s for silicon and gallium arsenide.

The electrons in the semiconductor are therefore moving rapidly in all directions. The thermal motion of an individual electron may be visualized as a succession of random scattering from collisions with lattice atoms, impurity atoms, and other scattering centers, as illustrated in Fig. 1a. The random motion of electrons leads to zero net displacement of an electron over a sufficiently long period of time. The average distance between collisions is called the *mean free path*, and the average time between collisions is called the *mean free time* τ_c . For a typical value of 10^{-5} cm for the mean free path, τ_c is about 1 ps (i.e., $10^{-5}/v_{th} \cong 10^{-12}$ s).

When a small electric field \mathcal{E} is applied to the semiconductor sample, each electron will experience a force $-q\mathcal{E}$ from the field and will be accelerated along the field (in the opposite direction to the field) during the time between collisions. Therefore, an additional velocity component will be superimposed upon the thermal motion of electrons. This additional component is called the *drift velocity*. The combined displacement of an electron due to the random thermal motion and the drift component is illustrated in Fig. 1b. Note that there is a net displacement of the electron in the direction opposite to the applied field.

We can obtain the drift velocity v_n by equating the momentum (force \times time) applied to an electron during the free flight between collisions to the momentum gained by the electron in the same period. The equality is valid because in a steady state all momentum gained between collisions is lost to the lattice in the collision. The momentum applied to an electron is given by $-q\mathcal{E}\tau_c$, and the momentum gained is $m_n v_n$. We have

$$q\mathcal{E}\tau_c = m_n v_n \quad (2)$$

or

$$v_n = \left(\frac{q\tau_c}{m_n} \right) \mathcal{E}. \quad (2a)$$

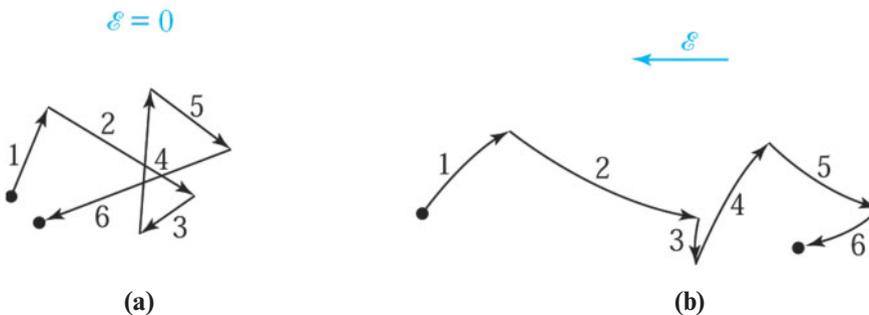


Fig. 1 Schematic path of an electron in a semiconductor. (a) Random thermal motion. (b) Combined motion due to random thermal motion and an applied electric field.

Equation 2a states that the electron drift velocity is proportional to the applied electric field. The proportionality factor depends on the mean free time and the effective mass. The proportionality factor is called the *electron mobility* μ_n with units of $\text{cm}^2/\text{V}\cdot\text{s}$, or

$$\mu_n \equiv \frac{q\tau_c}{m_n}. \quad (3)$$

Thus,

$$v_n = \mu_n \mathcal{E}. \quad (4)$$

Mobility is an important parameter for carrier transport because it describes how strongly the motion of an electron is influenced by an applied electric field. A similar expression can be written for holes in the valence band:

$$v_p = -\mu_p \mathcal{E}, \quad (5)$$

where v_p is the hole drift velocity and μ_p is the hole mobility. The negative sign is removed in Eq. 5, because holes drift in the same direction as the electric field.

In Eq. 3 the mobility is related directly to the mean free time between collisions, which in turn is determined by the various scattering mechanisms. The two most important mechanisms are lattice scattering and impurity scattering. Lattice scattering results from thermal vibrations of the lattice atoms at any temperature above absolute zero. These vibrations disturb the lattice periodic potential and allow energy to be transferred between the carriers and the lattice. Since lattice vibration increases with increasing temperature, lattice scattering becomes dominant at high temperatures; hence the mobility decreases with increasing temperature. Theoretical analysis¹ shows that the mobility due to lattice scattering μ_L will decrease in proportion to $T^{-3/2}$.

Impurity scattering results when a charge carrier travels past an ionized dopant impurity (donor or acceptor). The charge carrier path will be deflected because of Coulomb force interaction. The probability of impurity scattering depends on the total concentration of ionized impurities, that is, the sum of the concentration of negatively and positively charged ions. However, unlike lattice scattering, impurity scattering becomes less significant at higher temperatures. At higher temperatures, the carriers move faster; they remain near the impurity atom for a shorter time and are therefore scattered less effectively. The mobility due to impurity scattering μ_I can be shown to vary as $T^{3/2}/N_T$, where N_T is the total impurity concentration.²

The number of collisions taking place in a unit time, $1/\tau_c$, is the sum of the numbers of collisions due to the various scattering mechanisms:

$$\frac{1}{\tau_c} = \frac{1}{\tau_{c, \text{lattice}}} + \frac{1}{\tau_{c, \text{impurity}}}, \quad (6)$$

or

$$\frac{1}{\mu} = \frac{1}{\mu_L} + \frac{1}{\mu_I} \quad (6a)$$

Figure 2 shows the measured electron mobility as a function of temperature for silicon with five different donor concentrations.³ The inset shows the theoretical temperature dependence of mobility due to both lattice and impurity scatterings. For lightly doped samples (e.g., the sample with doping of 10^{14} cm^{-3}), the lattice scattering dominates, and the mobility decreases as the temperature increases. For heavily doped samples, the effect of impurity scattering is most pronounced at low temperatures. The mobility increases as the temperature increases, as can be seen for the sample with doping of 10^{19} cm^{-3} . For a given temperature, the mobility decreases with increasing impurity concentration because of enhanced impurity scatterings.

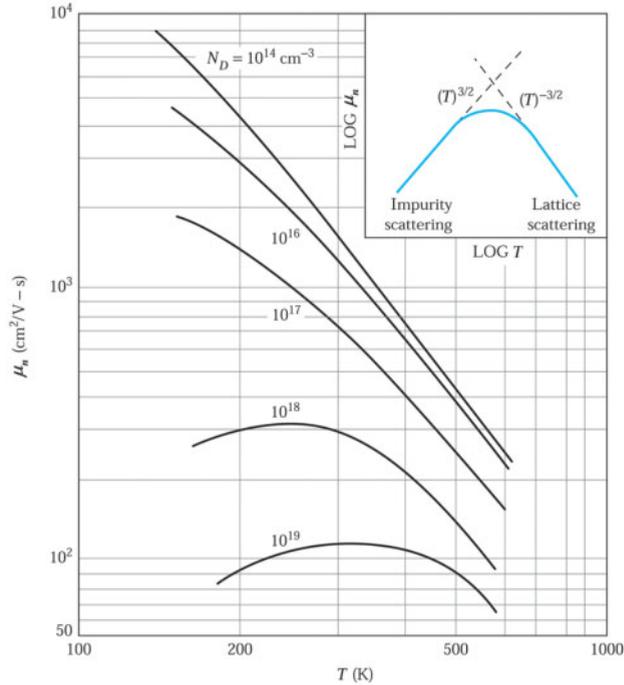


Fig. 2 Electron mobility in silicon versus temperature for various donor concentrations. Inset shows the theoretical temperature dependence of electron mobility.

Figure 3 shows the measured mobilities and diffusivities in silicon and gallium arsenide as a function of impurity concentration at room temperature.³ Mobility reaches a maximum value at low impurity concentrations; this corresponds to the lattice-scattering limitation. Both electron and hole mobilities decrease with increasing impurity concentration and eventually approach a minimum value at high impurity concentrations. Note also that the mobility of electrons is greater than that of holes. Greater electron mobility is due mainly to the smaller effective mass of electrons.

► EXAMPLE 1

Calculate the mean free time of an electron having a mobility of $1000 \text{ cm}^2/\text{V}\cdot\text{s}$ at 300 K ; also calculate the mean free path. Assume $m_n = 0.26 m_0$ in these calculations.

SOLUTION From Eq. 3, the mean free time is given by

$$\begin{aligned}\tau_c &= \frac{m_n \mu_n}{q} = \frac{(0.26 \times 0.91 \times 10^{-30} \text{ kg}) \times (1000 \times 10^{-4} \text{ m}^2 / \text{V}\cdot\text{s})}{1.6 \times 10^{-19} \text{ C}} \\ &= 1.48 \times 10^{-13} \text{ s} = 0.148 \text{ ps}.\end{aligned}$$

The thermal velocity is $2.28 \times 10^7 \text{ cm/s}$ for $m_n = 0.26 m_0$ from Eq. (1).

The mean free path is given by

$$l = v_{th} \tau_c = (3kT/m_n)^{1/2} \tau_c = (2.28 \times 10^7 \text{ cm/s})(1.48 \times 10^{-13} \text{ s}) = 3.37 \times 10^{-6} \text{ cm} = 33.7 \text{ nm}.$$

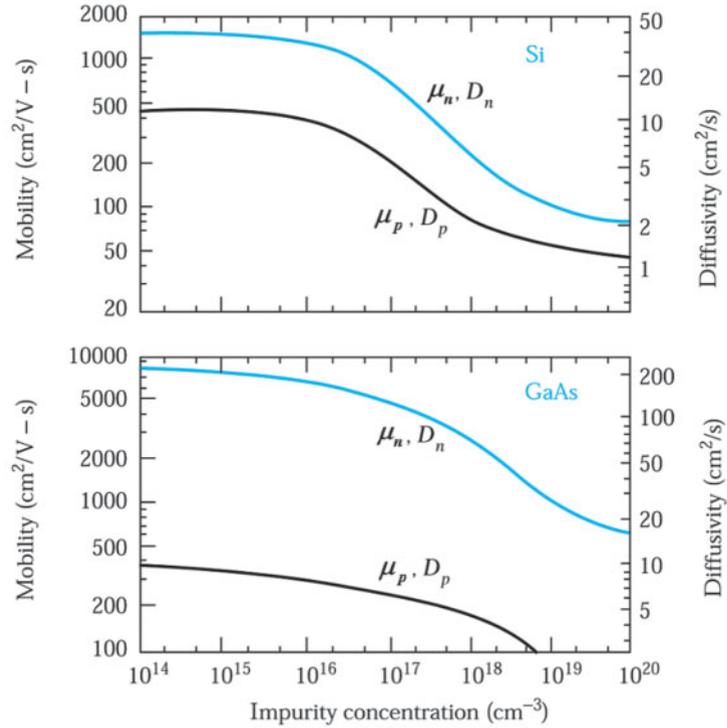


Fig. 3 Mobilities and diffusivities in Si and GaAs at 300 K as a function of impurity concentration³.

2.1.2 Resistivity

We now consider conduction in a homogeneous semiconductor material. Figure 4a shows an n -type semiconductor and its band diagram at thermal equilibrium. Figure 4b shows the corresponding band diagram when a positive biasing voltage is applied to the right-hand terminal. We assume that the contacts at the left-hand and right-hand terminals are ohmic, that is, there is negligible voltage drop at each of the contacts. The behavior of ohmic contacts is considered in Chapter 7. As mentioned previously, when an electric field \mathcal{E} is applied to a semiconductor, each electron will experience a force $-q\mathcal{E}$ from the field. The force is equal to the negative gradient of potential energy; that is,

$$-q\mathcal{E} = -(\text{gradient of electron potential energy}) = \frac{dE_C}{dx}. \quad (7)$$

Recall that in Chapter 1, the bottom of the conduction band E_C corresponds to the potential energy of an electron. Since we are interested in the gradient of the potential energy, we can use any part of the band diagram that is parallel to E_C (e.g., E_F , E_p or E_V , as shown in Fig. 4b). It is convenient to use the intrinsic Fermi level E_i because we shall use E_i when we consider p - n junctions in Chapter 3. Therefore, from Eq. 7 we have

$$\mathcal{E} = \frac{1}{q} \frac{dE_C}{dx} = \frac{1}{q} \frac{dE_i}{dx}. \quad (8)$$

We can define a related quantity ψ as the *electrostatic potential* whose negative gradient equals the electric field:

$$\mathcal{E} \equiv -\frac{d\psi}{dx}. \quad (9)$$

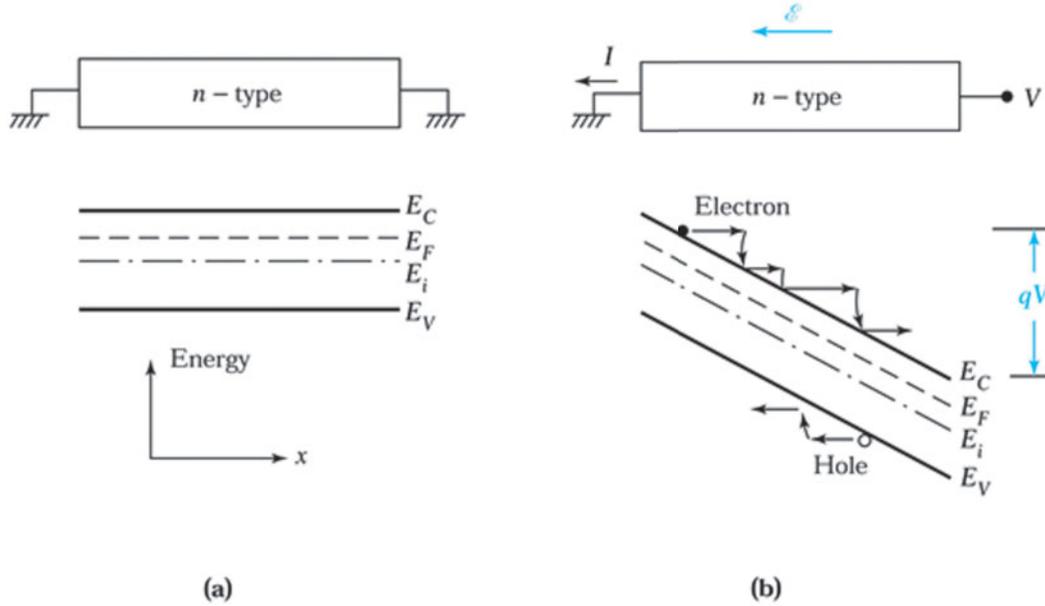


Fig. 4 Conduction process in an n-type semiconductor (a) at thermal equilibrium and (b) under a biasing condition.

Comparison of Eqs. 8 and 9 gives

$$\psi = -\frac{E_i}{q}, \tag{10}$$

which provides a relationship between the electrostatic potential and the potential energy of an electron. For the homogeneous semiconductor shown in Fig. 4b, the potential energy and E_i decrease linearly with distance; thus, the electric field is a constant in the negative x -direction. Its magnitude is the applied voltage divided by the sample length.

The electrons in the conduction band move to the right side, as shown in Fig. 4b. The kinetic energy corresponds to the distance from the band edge (i.e., E_C for electrons). When an electron undergoes a collision, it loses some or all of its kinetic energy to the lattice and drops toward its thermal equilibrium position. After the electron has lost some or all its kinetic energy, it will again begin to move toward the right and the same process will be repeated many times. Conduction by holes can be visualized in a similar manner but in the opposite direction.

The transport of carriers under the influence of an applied electric field produces a current called the drift current. Consider a semiconductor sample shown in Fig. 5, that has a cross-sectional area A , a length L , and a carrier concentration of n electrons/cm³. Suppose we now apply an electric field \mathcal{E} to the sample. The electron current density J_n flowing in the sample can be found by summing the product of the charge ($-q$) on each electron and the electron velocity over all electrons (n) per unit volume.

$$J_n = \frac{I_n}{A} = \sum_{i=1}^n (-qv_i) = -qnv_n = -qn\mu_n\mathcal{E}, \tag{11}$$

where I_n is the electron current. We have employed Eq. 4 for the relationship between v_n and \mathcal{E} . A similar argument applies to holes. By taking the charge on the hole to be positive, we have

$$J_p = qp v_p = qp\mu_p\mathcal{E}. \tag{12}$$

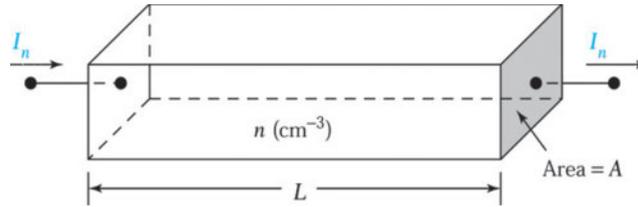


Fig. 5 Current conduction in a uniformly doped semiconductor bar with length L and cross-sectional area A .

The total current flowing in the semiconductor sample due to the applied field \mathcal{E} can be written as the sum of the electron and hole current components:

$$J = J_n + J_p = (qn\mu_n + qp\mu_p)\mathcal{E}. \quad (13)$$

The quantity in parentheses is known as *conductivity*:

$$\sigma = q(n\mu_n + p\mu_p). \quad (14)$$

The electron and hole contributions to conductivity are simply additive.

The corresponding resistivity of the semiconductor, which is the reciprocal of σ , is given by

$$\rho \equiv \frac{1}{\sigma} = \frac{1}{q(n\mu_n + p\mu_p)}. \quad (15)$$

Generally, in extrinsic semiconductors, only one of the components in Eq. 13 or 14 is significant because of the many orders-of-magnitude difference between the two carrier densities. Therefore, Eq. 15 reduces to

$$\rho = \frac{1}{qn\mu_n} \quad (15a)$$

for an n -type semiconductor (since $n \gg p$), and to

$$\rho = \frac{1}{qp\mu_p} \quad (15b)$$

for a p -type semiconductor (since $p \gg n$).

The most common method for measuring resistivity is the four-point probe method shown in Fig. 6. The probes are equally spaced. A small current I from a constant-current source is passed through the outer two probes and a voltage V is measured between the inner two probes. For a thin semiconductor sample with thickness W that is much smaller than the diameter d , the resistivity is given by

$$\rho = \frac{V}{I} \cdot W \cdot CF \quad \Omega\text{-cm}, \quad (16)$$

where CF is a well-documented “correction factor”. The correction factor depends on the ratio of d/s , where s is the probe spacing. When $d/s > 20$, the correction factor approaches 4.54. Figure 7 shows the measured room-temperature resistivity as a function of the impurity concentration for silicon and gallium arsenide. At this temperature and for low impurity concentrations, all donor (e.g., P and As in Si) or acceptor (e.g., B in Si) impurities that have shallow energy levels will be ionized. Under these conditions, the carrier concentration is equal to the impurity concentration. From these curves we can obtain the impurity concentration of a semiconductor if the resistivity is known, or vice versa.

► EXAMPLE 2

Find the room-temperature resistivity of an n -type silicon doped with 10^{16} phosphorus atoms/cm³.

SOLUTION At room temperature we assume that all donors are ionized; thus,

$$n \approx N_D = 10^{16} \text{ cm}^{-3}.$$

From Fig. 7 we find $\rho = 0.5 \text{ } \Omega\text{-cm}$. We can also calculate the resistivity from Eq. 15a:

$$\rho = \frac{1}{qn\mu_n} = \frac{1}{1.6 \times 10^{-19} \times 10^{16} \times 1300} = 0.48 \text{ } \Omega\text{-cm}.$$

The mobility μ_n is obtained from Fig. 3.

2.1.3 The Hall Effect

The carrier concentration in a semiconductor may be different from the impurity concentration, because the ionized impurity density depends on the temperature and the impurity energy level. To measure the carrier concentration directly, the most common method is the Hall effect. Hall measurement is also one of the most convincing methods to show the existence of holes as charge carriers, because the measurement can

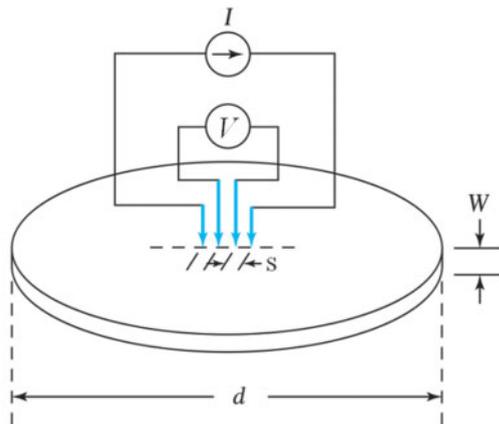


Fig. 6 Measurement of resistivity using a four-point probe.³

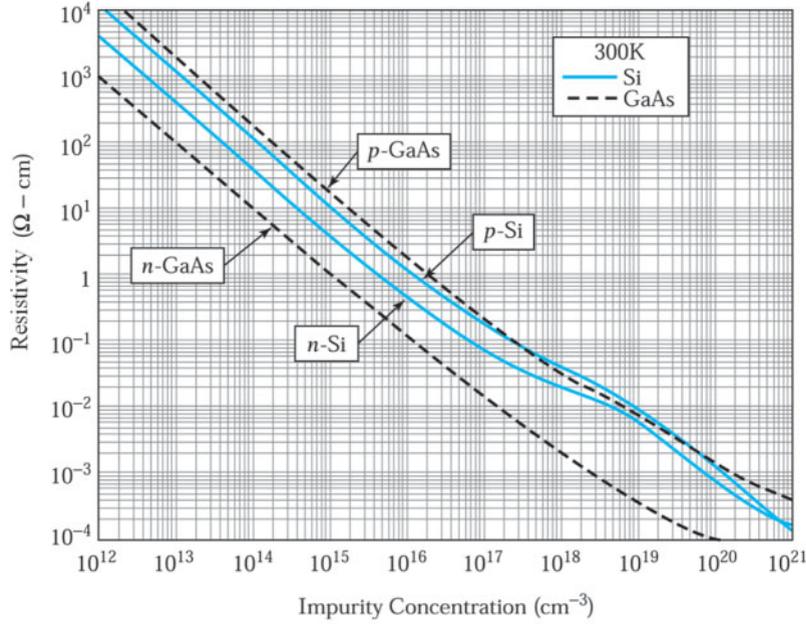


Fig. 7 Resistivity versus impurity concentration³ for Si and GaAs.

give the carrier type directly. Figure 8 shows an electric field applied along the x -axis and a magnetic field applied along the z -axis. Consider a p -type semiconductor sample. The Lorentz force $q\mathbf{v} \times \mathbf{B}$ ($= qv_x B_z$) due to the magnetic field will exert an average upward force on the holes flowing in the x -direction. The upward Lorentz force causes an accumulation of holes at the top of the sample that gives rise to a downward-directed electric field \mathcal{E}_y . Since there is no net current flow along the y -direction in the steady state, the electric field along the y -axis exactly balances the Lorentz force; that is,

$$q\mathcal{E}_y = qv_x B_z, \quad (17)$$

or

$$\mathcal{E}_y = v_x B_z. \quad (18)$$

Once the electric field \mathcal{E}_y becomes equal to $v_x B_z$, holes do not experience a net force along the y -direction as they drift in the x -direction.

The establishment of the electric field is known as the *Hall effect*. The electric field in Eq. 18 is called the *Hall field*, and the terminal voltage $V_H = \mathcal{E}_y W$ (Fig. 8) is called the *Hall-voltage*. Using Eq. 12 for the hole drift velocity, the Hall field \mathcal{E}_y in Eq. 18 becomes

$$\mathcal{E}_y = \left[\frac{J_x}{qp} \right] B_z = R_H J_x B_z, \quad (19)$$

where

$$R_H \equiv \frac{1}{qp}. \quad (20)$$

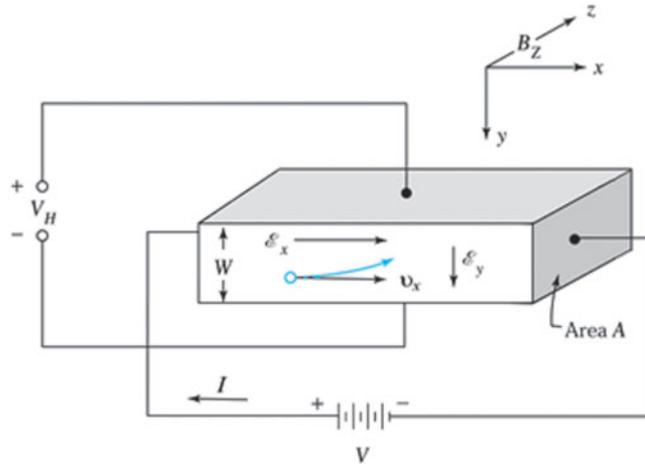


Fig. 8 Basic setup to measure carrier concentration using the Hall effect.

The Hall field \mathcal{E}_y is proportional to the product of the current density and the magnetic field. The proportionality constant R_H is the *Hall coefficient*. A similar result can be obtained for an *n*-type semiconductor, except that the Hall coefficient is negative:

$$R_H = -\frac{1}{qn}. \quad (21)$$

A measurement of the Hall voltage for a known current and magnetic field yields

$$p = \frac{1}{qR_H} = \frac{J_p B_z}{q\mathcal{E}_y} = \frac{(I/A)B_z}{q(V_H/W)} = \frac{IB_z W}{qV_H A}, \quad (22)$$

where all the quantities in the right-hand side of the equation can be measured. Therefore, the carrier concentration and carrier type can be obtained directly from the Hall measurement.

► EXAMPLE 3

A sample of Si is doped with 10^{16} phosphorus atoms/cm³. Find the Hall voltage in a sample with $W = 500\mu\text{m}$, $A = 2.5 \times 10^{-3} \text{ cm}^2$, $I = 1 \text{ mA}$, and $B_z = 10^{-4} \text{ Wb/cm}^2$.

SOLUTION The Hall coefficient is

$$R_H = \frac{1}{qn} = \frac{1}{1.6 \times 10^{19} \times 10^{16}} = 625 \text{ cm}^3/\text{C}$$

The Hall voltage is

$$\begin{aligned} V_H &= \mathcal{E}_y W = \left[R_H \frac{I}{A} B_z \right] W \\ &= \left[625 \cdot \frac{10^{-3}}{2.5 \times 10^{-3}} \cdot 10^{-4} \right] 500 \times 10^{-4} \\ &= -1.25 \text{ mV}. \end{aligned}$$

► 2.2 CARRIER DIFFUSION

2.2.1 Diffusion Process

In the preceding section, we considered the drift current, that is, the transport of carriers when an electric field is applied. Another important current component can exist if there is a spatial variation of carrier concentration in a semiconductor material. The carriers tend to move from a region of high concentration to a region of low concentration. This current component is called the *diffusion current*.

To understand the diffusion process, let us assume an electron density that varies in the x -direction, as shown in Fig. 9. The semiconductor is at uniform temperature, so that the average thermal energy of electrons does not vary with x ; only the density $n(x)$ varies. Consider the number of electrons crossing the plane at $x = 0$ per unit time and per unit area. Because of finite temperature, the electrons have random thermal motions with a thermal velocity v_{th} and a mean free path l (note that v_{th} is the thermal velocity in the x -direction, $l = v_{th}\tau_c$, where τ_c is the mean free time). The electrons at $x = -l$, one mean free path away on the left side, have equal chances of moving left or right; and in a mean free time τ_c , one half of them will move across the plane $x = 0$. The average rate of electron flow per unit area F_1 of electrons crossing plane $x = 0$ from the left is then

$$F_1 = \frac{1}{2} \frac{n(-l) \cdot l}{\tau_c} = \frac{1}{2} n(-l) \cdot v_{th} . \quad (23)$$

Similarly, the average rate of electron flow per unit area F_2 of electrons at $x = l$ crossing plane $x = 0$ from the right is

$$F_2 = \frac{1}{2} n(l) \cdot v_{th} . \quad (24)$$

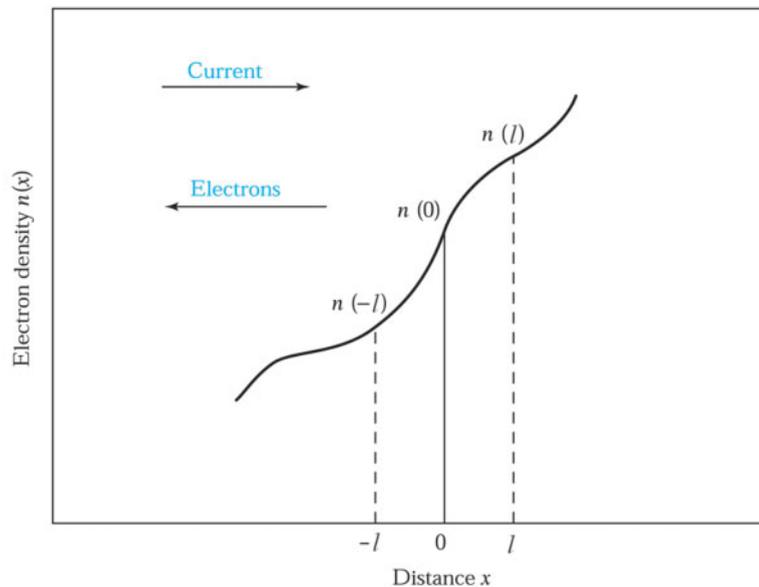


Fig. 9 Electron concentration versus distance; l is the mean free path. Arrows show the directions of electron and current flows.

The net rate of carrier flow from left to right is

$$F = F_1 - F_2 = \frac{1}{2}v_{th}[n(-l) - n(l)] \quad (25)$$

Approximating the densities at $x = \pm l$ by the first two terms of a Taylor series expansion, we obtain

$$F = \frac{1}{2}v_{th} \left\{ \left[n(0) - l \frac{dn}{dx} \right] - \left[n(0) + l \frac{dn}{dx} \right] \right\}$$

$$v_{th}l \frac{dn}{dx} \equiv D_n \frac{dn}{dx}, \quad (26)$$

where $D_n \equiv v_{th}l$ is called the *diffusion coefficient* or the *diffusivity*. Because each electron carries a charge $-q$, the carrier flow gives rise to a current

$$J_n = -qF = qD_n \frac{dn}{dx}. \quad (27)$$

The diffusion current is proportional to the spatial derivative of the electron density. Diffusion current results from the random thermal motion of carriers in a concentration gradient. For an electron density that increases with x , the gradient is positive and the electrons will diffuse toward the negative x -direction. The current is positive and flows in the direction opposite to that of the electrons, as indicated in Fig. 9.

► EXAMPLE 4

Assume that, in an n -type semiconductor at $T = 300\text{K}$, the electron concentration varies linearly from 1×10^{18} to $7 \times 10^{17} \text{ cm}^{-3}$ over a distance of 0.1 cm . Calculate the diffusion current density if the electron diffusion coefficient is $D_n = 22.5 \text{ cm}^2/\text{s}$.

SOLUTION The diffusion current density is given by

$$J_{n,diff} = qD_n \frac{dn}{dx} \approx qD_n \frac{\Delta n}{\Delta x}$$

$$(1.6 \times 10^{-19})(22.5) \left[\frac{1 \times 10^{18} - 7 \times 10^{17}}{0.1} \right] = 10.8 \text{ A/cm}^2. \quad \blacktriangleleft$$

2.2.2 Einstein Relation

Equation 27 can be written in a more useful form using the theorem for the equipartition of energy for this one-dimensional case. We can write

$$\frac{1}{2} m_n v_{th}^2 = \frac{1}{2} kT. \quad (28)$$

From Eqs. 3, 26, and 28 and using the relationship $l = v_{th}\tau_c$, we obtain

$$D_n = v_{th}l = v_{th}(v_{th}\tau_c) = v_{th}^2 \left(\frac{\mu_n m_n}{q} \right) = \left(\frac{kT}{m_n} \right) \left(\frac{\mu_n m_n}{q} \right), \quad (29)$$

or

$$D_n = \left(\frac{kT}{q} \right) \mu_n. \quad (30)$$

Equation 30 is known as the *Einstein relation*. It relates the two important constants (diffusivity and mobility) that characterize carrier transport by diffusion and by drift in a semiconductor. The Einstein relation also applies between D_p and μ_p . Values of diffusivities for silicon and gallium arsenide are shown in Fig. 3.

► EXAMPLE 5

Minority carriers (holes) are injected into a homogeneous n -type semiconductor sample at one point. An electric field of 50 V/cm is applied across the sample, and the field moves these minority carriers a distance of 1 cm in 100 μ s. Find the drift velocity and the diffusivity of the minority carriers. The temperature is 300 K.

SOLUTION

$$v_p = \frac{1 \text{ cm}}{100 \times 10^{-6} \text{ s}} = 10^4 \text{ cm/s};$$

$$\mu_p = \frac{v_p}{\mathcal{E}} = \frac{10^4}{50} = 200 \text{ cm}^2/\text{V}\cdot\text{s};$$

$$D_p = \frac{kT}{q} \mu_p = 0.0259 \times 200 = 5.18 \text{ cm}^2/\text{s}$$

2.2.3 Current Density Equations

When an electric field is present in addition to a concentration gradient, both drift current and diffusion current will flow. The total current density at any point is the sum of the drift and diffusion components:

$$J_n = q\mu_n n\mathcal{E} + qD_n \frac{dn}{dx}, \quad (31)$$

where \mathcal{E} is the electric field in the x -direction.

A similar expression can be obtained for the hole current:

$$J_p = q\mu_p p\mathcal{E} - qD_p \frac{dp}{dx}. \quad (32)$$

We use the negative sign in Eq. 32 because for a positive hole gradient the holes will diffuse in the negative x -direction. This diffusion results in a hole current that also flows in the negative x -direction. The total conduction current density is given by the sum of Eqs. 31 and 32:

$$J_{cond} = J_n + J_p. \quad (33)$$

The three expressions (Eqs. 31–33) constitute the current density equations. These equations are important for analyzing device operations under low electric fields. However, at sufficiently high electric fields the terms $\mu_n\mathcal{E}$ and $\mu_p\mathcal{E}$ should be replaced by the saturation velocity v_s discussed in Section 2.7.

► 2.3 GENERATION AND RECOMBINATION PROCESSES

In thermal equilibrium the relationship $pn = n_i^2$ is valid. If excess carriers are introduced to a semiconductor so that $pn > n_i^2$, we have a *nonequilibrium situation*. The process of introducing excess carriers is called *carrier injection*. Most semiconductor devices operate by the creation of charge carriers in excess of the thermal equilibrium values. We can introduce excess carriers by optical excitation or forward-biasing a p - n junction (discussed in Chapter 3).

Whenever the thermal-equilibrium condition is disturbed (i.e., $pn \neq n_i^2$), processes exist to restore the system to equilibrium (i.e., $pn = n_i^2$). In the case of injection of excess carriers, the mechanism that restores equilibrium is recombination of the injected minority carriers with the majority carriers. Depending on the nature of the recombination process, the energy released from the recombination process can be emitted as a photon or dissipated as heat to the lattice. When a photon is emitted, the process is called radiative recombination; otherwise, it is called nonradiative recombination.

Recombination phenomena can be classified as direct and indirect processes. Direct recombination, also called band-to-band recombination, usually dominates in direct-bandgap semiconductors, such as gallium arsenide, whereas indirect recombination via bandgap recombination centers dominates in indirect-bandgap semiconductors, such as silicon.

2.3.1 Direct Recombination

Consider a direct-bandgap semiconductor, such as GaAs, in thermal equilibrium. In terms of the band diagram, the thermal energy enables a valence electron to make an upward transition to the conduction band, leaving a hole in the valence band. This process is called carrier generation and is represented by the generation rate G_{th} (number of electron-hole pairs generated per cm^3 per second) in Fig. 10a. When an electron makes a transition downward from the conduction band to the valence band, an electron-hole pair is annihilated. This reverse process is called recombination; it is represented by the recombination rate R_{th} in Fig. 10a. Under thermal equilibrium conditions, the generation rate G_{th} must equal the recombination rate R_{th} , so that the carrier concentrations remain constant and the condition $pn = n_i^2$ is maintained.

When excess carriers are introduced to a direct-bandgap semiconductor, the probability is high that electrons and holes will recombine directly, because the bottom of the conduction band and the top of the valence band have the same momentum and no additional momentum is required for the transition across the bandgap. The rate of the direct recombination R is expected to be proportional to the number of electrons available in the conduction band and the number of holes available in the valence band; that is,

$$R = \beta np, \quad (34)$$

where β is the proportionality constant. As discussed previously, in thermal equilibrium the recombination rate must be balanced by the generation rate. Therefore, for an n -type semiconductor, we have

$$G_{th} = R_{th} = \beta n_{no} p_{no} \quad (35)$$

In this notation for carrier concentrations the first subscript refers to the type of the semiconductor. The subscript o indicates an equilibrium quantity. The n_{no} and p_{no} represent electron and hole densities, respectively, in an n -type semiconductor at thermal equilibrium. When we shine a light on the semiconductor to produce electron-hole pairs at a rate G_L (Fig. 10b), the carrier concentrations are above their equilibrium values. The recombination and generation rate become

$$R = \beta n_n p_n = \beta (n_{no} + \Delta n) (p_{no} + \Delta p), \quad (36)$$

$$G = G_L + G_{th}, \quad (37)$$

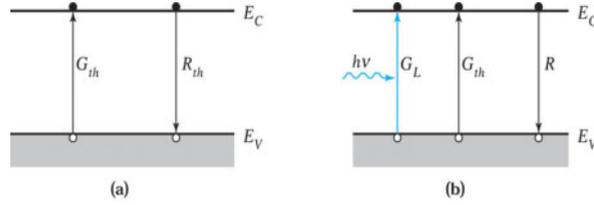


Fig. 10 Direct generation and recombination of electron-hole pairs: (a) at thermal equilibrium and (b) under illumination.

where Δn and Δp are the excess carrier concentrations, given by

$$\Delta n = n_n - n_{no} \quad (38a)$$

$$\Delta p = p_n - p_{no}, \quad (38b)$$

and $\Delta n = \Delta p$ to maintain overall charge neutrality.

The net rate of change of hole concentration is given by

$$\frac{dp_n}{dt} = G - R = G_L + G_{th} - R. \quad (39)$$

In steady state, $dp_n / dt = 0$. From Eq. 39 we have

$$G_L = R - G_{th} \equiv U, \quad (40)$$

where U is the net recombination rate. Substituting Eqs. 35 and 36 into Eq. 40 yields

$$U = \beta(n_{no} + p_{no} + \Delta p)\Delta p. \quad (41)$$

For low-level injection $\Delta p, p_{no} \ll n_{no}$, Eq. 41 is simplified to

$$U \cong \beta n_{no} \Delta p = \frac{p_n - p_{no}}{\frac{1}{\beta n_{no}}}. \quad (42)$$

Therefore, the net recombination rate is proportional to the excess minority carrier concentration. Obviously, $U = 0$ in thermal equilibrium. The proportionality constant $1/\beta n_{no}$ is called the *lifetime* τ_p of the excess minority carriers, or

$$U = \frac{p_n - p_{no}}{\tau_p}, \quad (43)$$

where

$$\tau_p \equiv \frac{1}{\beta n_{no}}. \quad (44)$$

The physical meaning of lifetime can best be illustrated by the transient response of a device after the sudden removal of the light source. Consider an n -type sample, as shown in Fig. 11a, that is illuminated

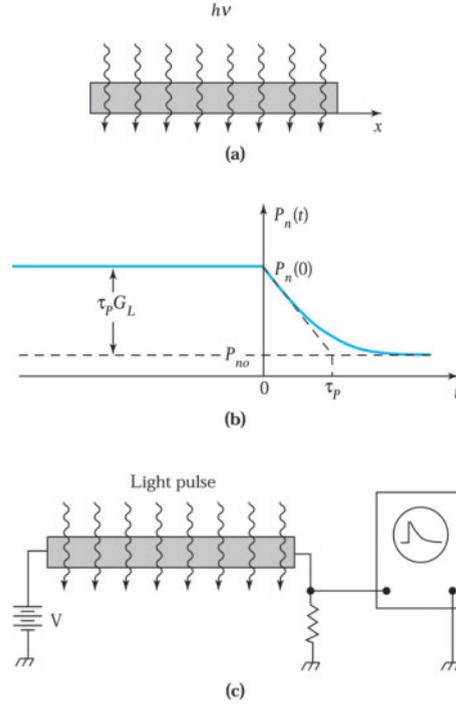


Fig. 11 Decay of photoexcited carriers. (a) n -type sample under constant illumination. (b) Decay of minority carriers (holes) with time. (c) Schematic setup to measure minority carrier lifetime.

with light and in which the electron-hole pairs are generated uniformly throughout the sample with a generation rate G_L . The time-dependent expression is given by Eq. 39. In steady state, from Eqs. 40 and 43

$$G_L = U = \frac{p_n - P_{no}}{\tau_p} \quad (45)$$

$$p_n = P_{no} + \tau_p G_L, \quad (45a)$$

$$\Delta n = \Delta p = \tau_p G_L. \quad (45b)$$

If at an arbitrary time, say $t = 0$, the light is suddenly turned off, the boundary condition is $p_n(t = 0) = P_{no} + \tau_p G_L$, as given by Eq. 45a. The time-dependent expression of Eq. 39 becomes

$$\frac{dp_n}{dt} = -\frac{p_n - P_{no}}{\tau_p} \quad (46)$$

and the solution is

$$p_n(t) = P_{no} + \tau_p G_L \exp(-t/\tau_p). \quad (47)$$

Figure 11b shows the variation of p_n with time. The minority carriers recombine with majority carriers and decay exponentially with a time constant τ_p , which corresponds to the lifetime defined in Eq. 44. Note that $p_n(t \rightarrow \infty) = P_{no}$.

This case illustrates the main idea of measuring the carrier lifetime using the photoconductivity method. Figure 11c shows a schematic setup. The excess carriers, generated uniformly throughout the sample by the light pulse, cause a momentary increase in the conductivity. The increase in conductivity manifests itself by a drop in voltage across the sample when a constant current is passed through it. The decay of the conductivity can be observed on an oscilloscope and is a measure of the lifetime of the excess minority carriers.

► EXAMPLE 6

A GaAs sample with $n_{no} = 10^{14} \text{ cm}^{-3}$ is illuminated with light and 10^{13} electron-hole pairs/ cm^3 are created every microsecond. If $\tau_n = \tau_p = 2 \mu\text{s}$, find the change in the minority carrier concentration.

SOLUTION Before illumination,

$$p_{no} = n_i^2 / n_{no} = (9.65 \times 10^9)^2 / 10^{14} \approx 9.31 \times 10^5 \text{ cm}^{-3}.$$

After illumination,

$$p_n = p_{no} + \tau_p G_L = 9.31 \times 10^5 + 2 \times 10^{-6} \times \frac{10^{13}}{1 \times 10^{-6}} \approx 2 \times 10^{13} \text{ cm}^{-3}.$$

$$\Delta p_n = \tau_p G_L = 2 \times 10^{13} \text{ cm}^{-3}.$$

2.3.2 Quasi-Fermi Level

Excess carriers are introduced to a semiconductor under the illumination of light. Electron and hole concentrations are higher than those in equilibrium state, such that $pn > n_i^2$. The Fermi level E_F is meaningful only in the thermal equilibrium state without any excess carriers. The quasi-Fermi levels E_{Fn} and E_{Fp} are used to express the electron and hole concentrations in nonequilibrium state and are defined by the following equations:

$$n = n_i e^{(E_{Fn} - E_i)/kT}, \quad (48)$$

$$p = n_i e^{(E_i - E_{Fp})/kT}. \quad (49)$$

► EXAMPLE 7

A GaAs sample with $n_{no} = 10^{16} \text{ cm}^{-3}$ is illuminated with light and 10^{13} electron-hole pairs/ cm^3 are created every microsecond. If $\tau_p = \tau_n = 2 \text{ ns}$, find the quasi-Fermi level at room temperature.

SOLUTION Before illumination

$$n_{no} = 10^{16} \text{ cm}^{-3};$$

$$p_{no} = n_i^2 / n_{no} = (2.25 \times 10^6)^2 / 10^{16} \approx 5.06 \times 10^{-4} \text{ cm}^{-3}.$$

The Fermi level measured from the intrinsic Fermi level is 0.575 eV.

After illumination, the electron and hole concentration are given:

$$n = n_{no} + \tau_n G_L = 10^{16} + 2 \times 10^{-9} \times \frac{10^{13}}{1 \times 10^{-6}} \approx 10^{16} \text{ cm}^{-3}$$

$$p = p_{no} + \tau_p G_L = 9.31 \times 10^3 + 2 \times 10^{-9} \times \frac{10^{13}}{1 \times 10^{-6}} \approx 2 \times 10^{10} \text{ cm}^{-3}$$

The quasi-Fermi levels at room temperature are given by Eq. 48 and 49:

$$E_{Fn} - E_i = kT \ln(n_n / n_i) = 0.0259 \ln(10^{16} / 2.25 \times 10^6) = 0.575 \text{ eV}$$

$$E_i - E_{Fp} = kT \ln(p_n / n_i) = 0.0259 \ln(2 \times 10^{10} / 2.25 \times 10^6) = 0.235 \text{ eV}$$

These results are shown in Fig. 12. ◀

From the example, it is obvious that the excitation causes a large percentage change in the minority carrier concentration and almost no change in the majority concentration. The separation of the quasi-Fermi levels is a direct measure of the deviation from equilibrium. It is very useful to visualize majority and minority carrier concentrations varying with position in devices.

2.3.3 Indirect Recombination

For indirect-bandgap semiconductors, such as silicon, a direct recombination process is very unlikely, because the electrons at the bottom of the conduction band have nonzero momentum with respect to the holes at the top of the valence band (see Chapter 1). A direct transition that conserves both energy and momentum is not possible without a simultaneous lattice interaction. Therefore the dominant recombination process in such semiconductors is indirect transition via localized energy states in the forbidden energy gap.⁴ These states act as stepping stones between the conduction band and the valence band.

Figure 13 shows various transitions that occur in the recombination process through intermediate-level states (also called recombination centers). We illustrate the charging state of the center before and after each of the four basic transitions taking place. The arrows in the figure designate the transition of the electron in a particular process. The illustration is for the case of a recombination center with a single energy level that is neutral when not occupied by an electron and negative when it is occupied. In indirect recombination, the derivation of the recombination rate is more complicated; the detailed derivation is given in Appendix I. The recombination rate is given by

$$U = \frac{\nu_{th} \sigma_n \sigma_p N_t (p_n n_n - n_i^2)}{\sigma_p [p_n + n_i e^{(E_i - E_i)/kT}] + \sigma_n [n_n + n_i e^{(E_i - E_i)/kT}]}, \quad (50)$$

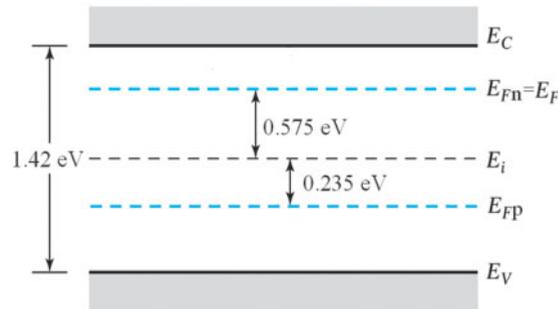


Fig. 12 Band diagram showing the quasi-Fermi levels.

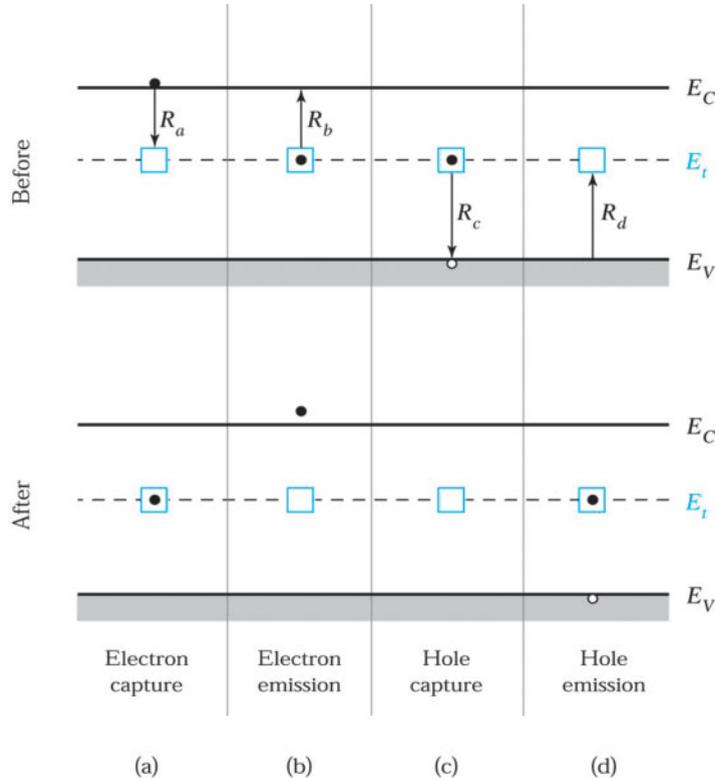


Fig. 13 Indirect generation-recombination processes of (a) electron capture, (b) electron emission, (c) hole capture, and (d) hole emission at thermal equilibrium.

where v_{th} is the thermal velocity of carriers given in Eq. 1, N_t is the concentration of the recombination center in the semiconductor, and σ_n is the electrons capture cross section. The quantity σ_n describes the effectiveness of the center in capturing an electron and is a measure of how close the electron has to come to the center to be captured. σ_p is the capture cross section of holes. E_t is the energy level of the recombination center.

We can simplify the general expression for the dependence of U on E_t by assuming equal electron and hole capture cross sections, that is, $\sigma_n = \sigma_p = \sigma_o$. Equation 50 then becomes

$$U = v_{th} \sigma_o N_t \frac{(p_n n_n - n_i^2)}{p_n + n_n + 2n_i \cosh\left(\frac{E_t - E_i}{kT}\right)}. \quad (51)$$

Under a low-injection condition in an n -type semiconductor $n_n \gg p_n$, the recombination rate can be written as

$$U \approx v_{th} \sigma_o N_t \frac{p_n - p_{no}}{1 + \left(\frac{2n_i}{n_{no}}\right) \cosh\left(\frac{E_t - E_i}{kT}\right)} = \frac{p_n - p_{no}}{\tau_p}. \quad (52)$$

The recombination rate for indirect recombination is given by the same expression as Eq. 43; however, τ_p depends on the locations of the recombination centers in the bandgap.

2.3.4 Surface Recombination

Figure 14 shows schematically the bonds at a semiconductor surface.⁵ Because of the abrupt discontinuity of the lattice structure at the surface, a large number of localized energy states or generation-recombination centers may be introduced at the surface region. These energy states, called *surface states*, may greatly enhance the recombination rate at the surface region. The kinetics of surface recombination is similar to those considered before for bulk centers. The total number of carriers recombining at the surface *per unit area* in unit time can be expressed in a form analogous to Eq. 50. For a low-injection condition, and for the limiting case where an electron concentration at the surface is essentially equal to the bulk majority carrier concentration, the total number of carriers recombining at the surface per unit area and unit time can be simplified to

$$U_s \cong v_{th} \sigma_p N_{st} (p_s - p_{no}), \quad (53)$$

where p_s denotes the hole concentrations at the surface, and N_{st} is the recombination center density per unit area in the surface region. Since the product $v_{th} \sigma_p N_{st}$ has its dimension in centimeters per second, it is called the *low-injection surface recombination velocity* S_{lr} :

$$S_{lr} \equiv v_{th} \sigma_p N_{st}. \quad (54)$$

► 2.4 CONTINUITY EQUATION

In the previous sections we considered individual effects such as drift due to an electric field, diffusion due to a concentration gradient, and recombination of carriers through intermediate-level recombination centers. We now consider the overall effect when drift, diffusion, and recombination occur simultaneously in a semiconductor material. The governing equation is called the *continuity equation*.

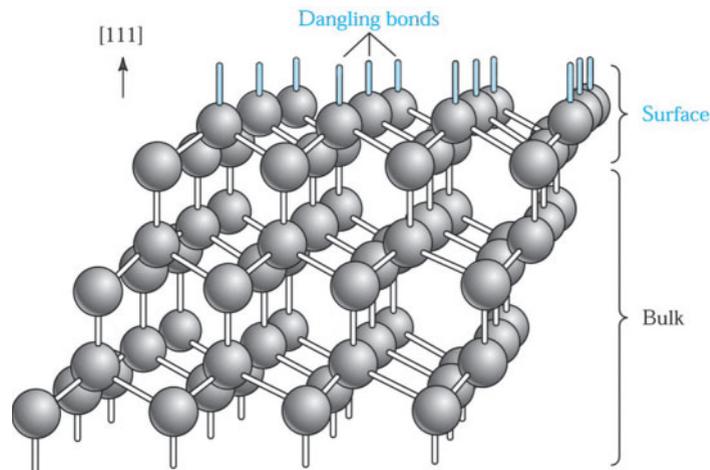


Fig. 14 Schematic diagram of bonds at a clean semiconductor surface. The bonds are anisotropic and differ from those in the bulk.⁵

To derive the one-dimensional continuity equation for electrons, consider an infinitesimal slice with thickness dx located at x , as shown in Fig. 15. The number of electrons in the slice may increase due to the net current flow into the slice and the *net* carrier generation in the slice. The overall rate of electron increase is the algebraic sum of four components: the number of electrons flowing into the slice at x , minus the number of electrons flowing out at $x + dx$, plus the rate at which electrons are generated, minus the rate at which they are recombined with holes in the slice.

The first two components are found by dividing the currents at each side of the slice by the charge of an electron. The generation and recombination rates are designated by G_n and R_n , respectively. The overall rate of change in the number of electrons in the slice is then

$$\frac{\partial n}{\partial t} A dx = \left[\frac{J_n(x)A}{-q} - \frac{J_n(x+dx)A}{-q} \right] + (G_n - R_n) A dx, \quad (55)$$

where A is the cross-sectional area and $A dx$ is the volume of the slice. Expanding the expression for the current at $x + dx$ in Taylor series yields

$$J_n(x+dx) = J_n(x) + \frac{\partial J_n}{\partial x} dx + \dots \quad (56)$$

We thus obtain the basic *continuity equation* for electrons:

$$\frac{\partial n}{\partial t} = \frac{1}{q} \frac{\partial J_n}{\partial x} + (G_n - R_n). \quad (57)$$

A similar continuity equation can be derived for holes, except that the sign of the first term on the right-hand side of Eq. 57 is changed because of the positive charge associated with a hole:

$$\frac{\partial p}{\partial t} = -\frac{1}{q} \frac{\partial J_p}{\partial x} + (G_p - R_p). \quad (58)$$

We can substitute the current expressions from Eqs. 31 and 32 and the recombination expressions from Eq. 43 into Eqs. 57 and 58. For the one-dimensional case under low-injection condition, the continuity equations for minority carriers (i.e., n_p in a p -type semiconductor or p_n in an n -type semiconductor) are

$$\boxed{\frac{\partial n_p}{\partial t} = n_p \mu_n \frac{\partial \mathcal{E}}{\partial x} + \mu_n \mathcal{E} \frac{\partial n_p}{\partial x} + D_n \frac{\partial^2 n_p}{\partial x^2} + G_n - \frac{n_p}{\tau_n} p_{po}}, \quad (59)$$

$$\boxed{\frac{\partial p_n}{\partial t} = p_n \mu_p \frac{\partial \mathcal{E}}{\partial x} + \mu_p \mathcal{E} \frac{\partial p_n}{\partial x} + D_p \frac{\partial^2 p_n}{\partial x^2} + G_p - \frac{p_n}{\tau_p} p_{no}}. \quad (60)$$

In addition to the continuity equations, Poisson's equation

$$\boxed{\frac{d\mathcal{E}}{dx} = \frac{\rho_s}{\epsilon_s}} \quad (61)$$

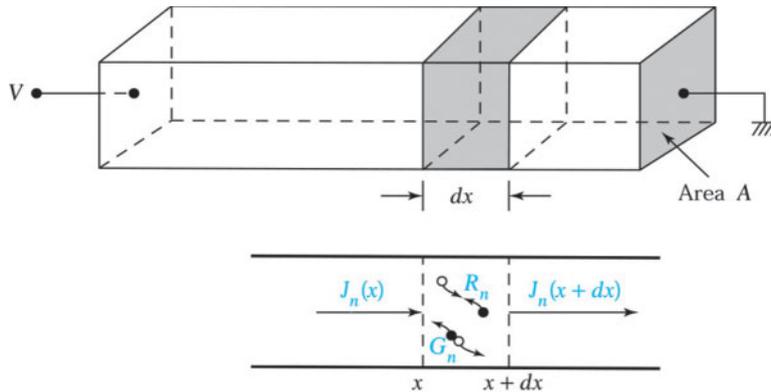


Fig. 15 Current flow and generation-recombination processes in an infinitesimal slice of thickness dx .

must be satisfied, where ϵ_s is the semiconductor dielectric permittivity and ρ_s is the space charge density given by the algebraic sum of the charge carrier densities and the ionized impurity concentrations, $q(p - n + N_D^+ - N_A^-)$.

In principle, Eqs. 59 through 61 together with appropriate boundary conditions have a unique solution. Because of the algebraic complexity of this set of equations, in most cases the equations are simplified with physical approximations before a solution is attempted. We solve the continuity equations for three important cases.

2.4.1 Steady-State Injection from One Side

Figure 16a shows an n -type semiconductor in which excess carriers are injected from one side as a result of illumination. It is assumed that light penetration is negligibly small (i.e., the assumptions of zero field and zero generation for $x > 0$). At steady state there is a concentration gradient near the surface. From Eq. 60

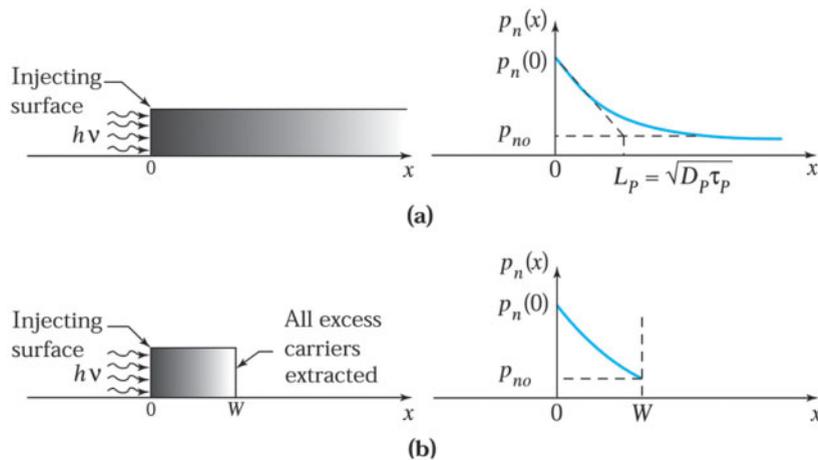


Fig. 16 Steady-state carrier injection from one side. (a) Semiinfinite sample. (b) Sample with thickness W .

the differential equation for the minority carriers inside the semiconductor is

$$\frac{\partial p_n}{\partial t} = 0 = D_p \frac{\partial^2 p_n}{\partial x^2} - \frac{p_n - p_{no}}{\tau_p}. \quad (62)$$

The boundary conditions are $p_n(x=0) = p_n(0) = \text{constant value}$ and $p_n(x \rightarrow \infty) = p_{no}$. The solution of $p_n(x)$ is

$$p_n(x) = p_{no} + [p_n(0) - p_{no}]e^{-x/L_p}. \quad (63)$$

The length L_p is equal to $\sqrt{D_p \tau_p}$, and is called the diffusion length. Figure 16a shows the variation of the minority carrier density, which decays with a characteristic length given by L_p .

If we change the second boundary condition as shown in Fig. 16b so that all excess carriers at $x = W$ are extracted, that is, $p_n(W) = p_{no}$, then we obtain a new solution for Eq. 62:

$$p_n(x) = p_{no} + [p_n(0) - p_{no}] \frac{\sinh\left(\frac{W-x}{L_p}\right)}{\sinh\left(\frac{W}{L_p}\right)}. \quad (64)$$

The current density at $x = W$ is given by the diffusion current expression, Eq. 32 with $\mathcal{E} = 0$:

$$J_p = -qD_p \left. \frac{\partial p_n}{\partial x} \right|_w = q[p_n(0) - p_{no}] \frac{D_p}{L_p} \frac{1}{\sinh(W/L_p)}. \quad (65)$$

2.4.2 Minority Carriers at the Surface

When surface recombination is introduced at one end of a semiconductor sample under illumination (Fig. 17), the hole current density flowing into the surface from the bulk of the semiconductor is given by qU_s . In this example, it is assumed that the sample is uniformly illuminated with uniform generation of carriers. The surface recombination leads to a lower carrier concentration at the surface. This gradient of hole concentration yields a diffusion current density that is equal to the surface recombination current. Therefore, the boundary condition at $x = 0$ is

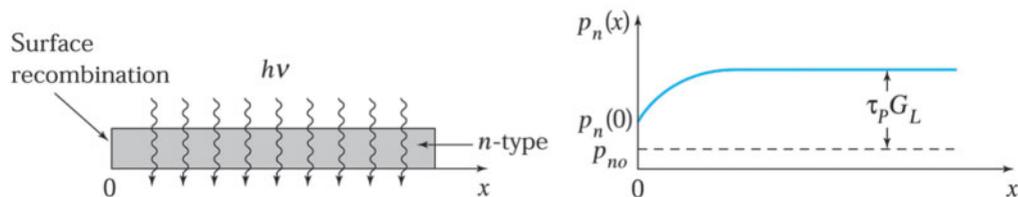


Fig. 17 Surface recombination at $x = 0$. The minority carrier distribution near the surface is affected by the surface recombination velocity.⁶

$$qD_p \left. \frac{dp_n}{dx} \right|_{x=0} = qU_s - qS_{lr}[p_n(0) - p_{no}]. \quad (66)$$

The boundary condition at $x = \infty$ is given by Eq. 45a. At steady state the differential equation is

$$\frac{\partial p_n}{\partial t} = 0 = D_p \frac{\partial^2 p_n}{\partial x^2} + G_L - \frac{p_n - p_{no}}{\tau_p}. \quad (67)$$

The solution of the equation, subject to the boundary conditions above, is⁶

$$p_n(x) = p_{no} + \tau_p G_L \left(1 - \frac{\tau_p S_{lr} e^{-x/L_p}}{L_p + \tau_p S_{lr}} \right). \quad (68)$$

A plot of this equation for a finite S_{lr} is shown in Fig. 17. When $S_{lr} \rightarrow 0$, then $p_n(x) \rightarrow p_{no} + \tau_p G_L$, which was obtained previously (Eq. 45a). When $S_{lr} \rightarrow \infty$, then

$$p_n(x) = p_{no} + \tau_p G_L (1 - e^{-x/L_p}). \quad (69)$$

From Eq. 69 we can see that at the surface the minority carrier density approaches its thermal equilibrium value p_{no} .

2.4.3 The Haynes-Shockley Experiment

One of the classic experiments in semiconductor physics is the demonstration of drift and diffusion of minority carriers, first made by Haynes and Shockley.⁷ The experiment allows independent measurement of the minority carrier mobility μ and diffusion coefficient D . The basic setup of the Haynes–Shockley experiment is shown in Fig. 18a. When localized light pulses generate excess minority carriers in a semiconductor, the transport equation after a pulse is given by Eq. 60 by setting $G_p = 0$ and $\partial \mathcal{E} / \partial x = 0$:

$$\frac{\partial p_n}{\partial t} = \mu_p \mathcal{E} \frac{\partial p_n}{\partial x} + D_p \frac{\partial^2 p_n}{\partial x^2} - \frac{p_n - p_{no}}{\tau_p}. \quad (70)$$

If no field is applied along the sample $\mathcal{E} = 0$, the solution is given by

$$p_n(x, t) = \frac{N}{\sqrt{4\pi D_p t}} \exp\left(-\frac{x^2}{4D_p t} - \frac{t}{\tau_p}\right) + p_{no}, \quad (71)$$

where N is the number of electrons or holes generated per unit area. Figure 18b shows this solution as the carriers diffuse away from the point of injection and also recombine.

If an electric field is applied along the sample, the solution is in the form of Eq. 71, but with x replaced by $x - \mu_p \mathcal{E} t$ (Fig. 18c); thus the whole “package” of excess minority carrier moves toward the negative end of the sample with the drift velocity $\mu_p \mathcal{E}$. At the same time, the carriers diffuse outward and recombine as in the field-free case. With a known sample length, applied oscilloscope, the drift field, and the time delay between the applied electric pulse and the detected pulse (both displayed on the mobility $\mu_p = L/\mathcal{E}t$ can be calculated.

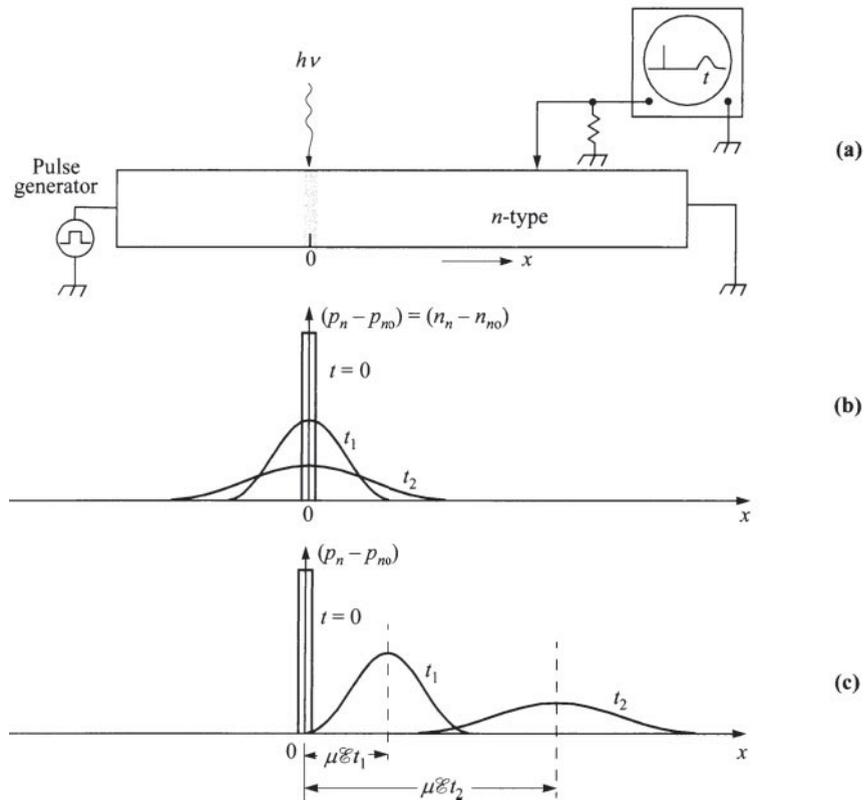


Fig. 18 The Haynes–Shockley experiment. (a) Experimental setup. (b) Carrier distributions without an applied field. (c) Carrier distributions with an applied field.⁷

▶ EXAMPLE 8

In a Haynes–Shockley experiment, the maximum amplitudes of the minority carriers at $t_1 = 100 \mu\text{s}$ and $t_2 = 200 \mu\text{s}$ differ by a factor of 5. Calculate the minority carrier lifetime.

SOLUTION When an electric field is applied, the minority carrier distribution is given by

$$\Delta p \equiv p_n - p_{no} = \frac{N}{\sqrt{4\pi D_p t}} \exp\left(-\frac{(x - \mu_p \mathcal{E} t)^2}{4D_p t} - \frac{t}{\tau_p}\right).$$

At the maximum amplitude

$$\Delta p = \frac{N}{\sqrt{4\pi D_p t}} \exp\left(-\frac{t}{\tau_p}\right).$$

Therefore

$$\frac{\Delta p(t_1)}{\Delta p(t_2)} = \frac{\sqrt{t_2} \exp(-t_1/\tau_p)}{\sqrt{t_1} \exp(-t_2/\tau_p)} = \frac{\sqrt{200}}{\sqrt{100}} \exp\left[\frac{200 - 100}{\tau_p (\mu\text{s})}\right] = 5$$

$$\therefore \tau_p = \frac{200 - 100}{\ln(5/\sqrt{2})} = 79 \mu\text{s}.$$

► 2.5 THERMIONIC EMISSION PROCESS

In previous sections, we considered carrier transport phenomena inside the bulk semiconductor. At the semiconductor surface, carriers may recombine via the recombination centers due to the dangling bonds at the surface region. In addition, if the carriers have sufficient energy, they may be “thermionically” emitted into the vacuum. This is called the *thermionic emission process*.

Figure 19a shows the band diagram of an isolated n -type semiconductor. The electron affinity, $q\chi$, is the energy difference between the conduction band edge and the vacuum level in the semiconductor; the work function, $q\phi_s$, is the energy between the Fermi level and the vacuum level in the semiconductor. From Fig. 19b, it is clear that an electron can be thermionically emitted into the vacuum if its energy is above $q\chi$.

The electron density with energies above $q\chi$ can be obtained from an expression similar to that for the electron density in the conduction band (Eqs. 6 and 13 of Chapter 1) except that the lower limit of the integration is $q\chi$ instead of E_C :

$$n_{th} = \int_{q\chi}^{\infty} n(E) dE = N_C \exp\left[-\frac{q(\chi + V_n)}{kT}\right], \quad (72)$$

where N_C is the effective density of states in the conduction band, and V_n is the difference between the bottom of the conduction band and the Fermi level.

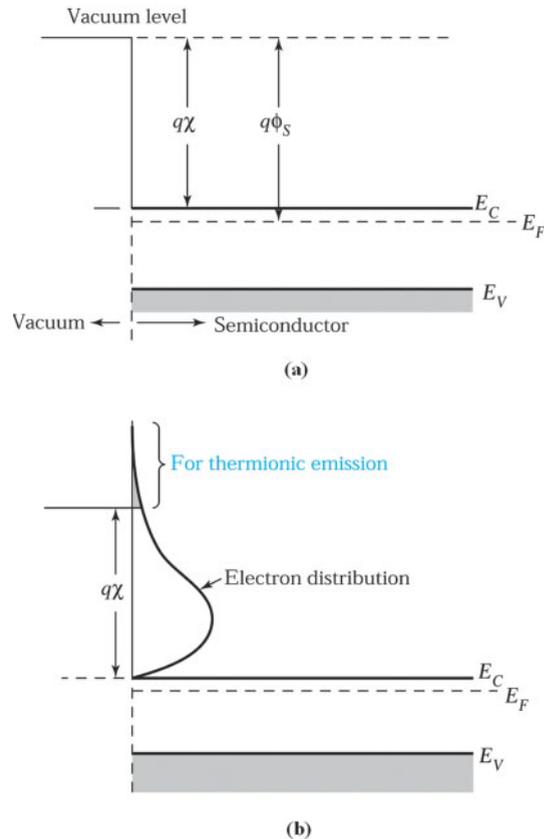


Fig. 19 (a) The band diagram of an isolated n -type semiconductor. (b) The thermionic emission process.

► EXAMPLE 9

Calculate the thermionically emitted electron density, n_{th} , at room temperature for an n -type silicon sample with an electron affinity of $q\chi = 4.05$ eV and $qV_n = 0.2$ eV. If we reduce the effective $q\chi$ to 0.6 eV, what is n_{th} ?

SOLUTION

$$n_{th}(4.05 \text{ eV}) = 2.86 \times 10^{19} \exp\left(\frac{4.05 + 0.2}{0.0259}\right) = 2.86 \times 10^{19} \exp(164) \cong 10^{52} \approx 0$$

$$n_{th}(0.6 \text{ eV}) = 2.86 \times 10^{19} \exp\left(\frac{0.8}{0.0259}\right) = 2.86 \times 10^{19} \exp(30.9) = 1 \times 10^6 \text{ cm}^{-3}$$

From the above example, we see that at 300 K there is no emission of electrons into vacuum for $q\chi = 4.05$ eV. However, if we can lower the effective electron affinity to 0.6 eV, a substantial number of electrons can be thermionically emitted. The thermionic emission process is of particular importance for metal-semiconductor contacts, to be considered in Chapter 7.

► 2.6 TUNNELING PROCESS

Figure 20a shows the energy band diagram when two isolated semiconductor samples are brought close together. The distance between them is d and the potential barrier height qV_0 is equal to the electron affinity $q\chi$. If the distance is sufficiently small, the electrons in the left-side semiconductor may transport across the barrier and move to the right-side semiconductor, even if the electron energy is much less than the barrier height. This process is associated with the *quantum tunneling phenomenon*.

Based on Fig. 20a, we have redrawn the one-dimensional potential barrier diagram in Fig. 20b. We first consider the transmission (or tunneling) coefficient of a particle (e.g., electron) through this barrier. In the corresponding classic case, the particle is always reflected if its energy E is less than the potential barrier height qV_0 . However, in the quantum case, the particle has finite probability to transmit or “tunnel” through the potential barrier.

The behavior of a particle (e.g., a conduction electron) in the region where $qV(x) = 0$ can be described by the Schrödinger equation:

$$\frac{\hbar^2}{2m_n} \frac{d^2\psi}{dx^2} = E\psi \quad (73)$$

or

$$\frac{d^2\psi}{dx^2} = \frac{2m_n E}{\hbar^2} \psi \quad (74)$$

where m_n is the effective mass, \hbar is the reduced Planck constant, E is the kinetic energy, and ψ is the wave function of the particle. The solutions are

$$\psi(x) = Ae^{jkx} + Be^{-jkx} \quad x \leq 0, \quad (75)$$

$$\psi(x) = Ce^{jkx} \quad x \geq d. \quad (76)$$

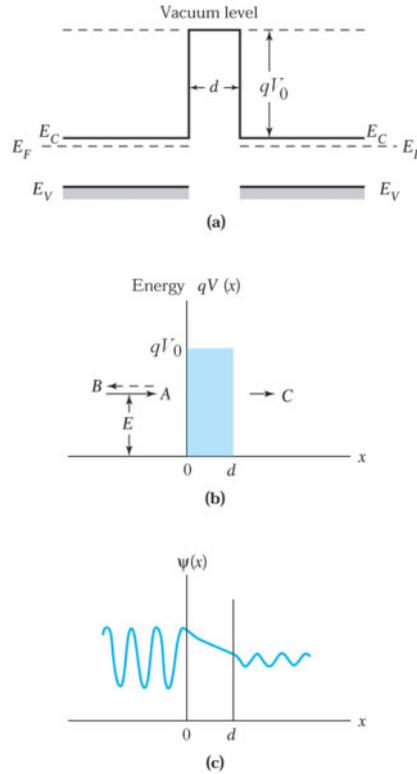


Fig. 20 (a) The band diagram of two isolated semiconductors separated by a distance d . (b) One-dimensional potential barrier. (c) Schematic representation of the wave function across the potential barrier.

where $k \equiv \sqrt{2m_n E / \hbar^2}$. For $x \leq 0$, we have an incident-particle wave function (with amplitude A) and a reflected wave function (with amplitude B); for $x \geq d$, we have a transmitted wave function (with amplitude C).

Inside the potential barrier, the wave equation is given by

$$\frac{\hbar^2}{2m_n} \frac{d^2\psi}{dx^2} + qV_0\psi = E\psi \quad (77)$$

or

$$\frac{d^2\psi}{dx^2} + \frac{2m_n(qV_0 - E)}{\hbar^2}\psi = 0 \quad (78)$$

The solution for $E < qV_0$ is

$$\psi(x) = Fe^{\beta x} + Ge^{-\beta x}, \quad (79)$$

where $\beta \equiv \sqrt{2m_n(qV_0 - E) / \hbar^2}$. A schematic representation of the wave functions across the barrier is shown in Fig. 20c. The continuity of ψ and $d\psi/dx$ at $x = 0$ and $x = d$, which is required by the boundary conditions, provides four relations between the five coefficients (A , B , C , F , and G). We can solve for $(C/A)^2$, which is the *transmission coefficient*:

$$\left(\frac{C}{A}\right)^2 = \left[1 + \frac{(qV_0 \sinh \beta d)^2}{4E(qV_0 - E)}\right]^{-1}. \quad (80)$$

The transmission coefficient decreases monotonically as E decreases. When $\beta d \gg 1$, the transmission coefficient becomes quite small and varies as

$$\boxed{\left[\frac{C}{A}\right]^2 \sim \exp(-2\beta d) = \exp\left[-2d\sqrt{2m_n(qV_0 - E)/\hbar^2}\right]}. \quad (81)$$

To have a finite transmission coefficient, we require a small tunneling distance d , a low-potential barrier qV_0 , and a small effective mass. These results will be used for tunnel diodes in Chapter 8.

► 2.7 SPACE-CHARGE EFFECT

The space charge in a semiconductor is determined by both the ionized impurity concentrations (N_D^+ and N_A^-) and the carrier concentrations (n and p),

$$\rho = q(p - n + N_D^+ - N_A^-) \quad (82)$$

In the neutral region of a semiconductor, $n = N_D^+$ and $p = N_A^-$, the space-charge density is zero. If we inject electrons into an n -type semiconductor (with $N_D^+ \gg N_A^- \approx p \approx 0$) so that the electron concentration n is much larger than N_D^+ , the space-charge density is no longer zero (i.e., $\rho \approx -qn$). The injected carrier density will effectively be the space-charge density, which will in turn determine the electric-field distribution from Poisson's equation. This is the space-charge effect.

In the presence of a space-charge effect, if the current is dominated by the drift component of the injected carriers, it is called the space-charge-limited current. Figure 21a shows the band diagram for the case of electron injection; the drift current is given by

$$J = qn v. \quad (83)$$

The space charge is determined by the injected carriers (assuming $n \gg N_D^+$, $p \approx N_A^- \approx 0$), giving rise to a Poisson's equation of the form

$$\frac{d\mathcal{E}}{dx} = \frac{\rho}{\epsilon_s} = \frac{qn}{\epsilon_s}. \quad (84)$$

In the constant-mobility regime,

$$v = \mu \mathcal{E}. \quad (85)$$

Substituting Eqs. 83 and 85 into Eq. 84 yields

$$\frac{d\mathcal{E}}{dx} = \frac{J}{\epsilon_s \mu \mathcal{E}}, \quad (86)$$

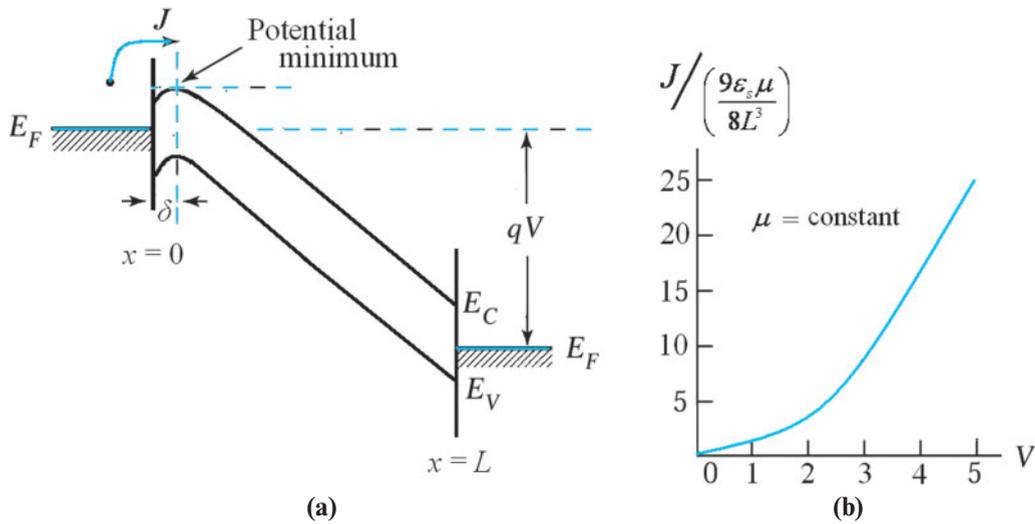


Fig. 21 Space-charge effect. (a) The band diagram for the case of electron injection. (b) The space-charge-limited current in the constant-mobility regime is proportional to the square of the applied voltage.

or

$$\mathcal{E} d\mathcal{E} = \frac{J}{\epsilon_s \mu} dx. \quad (87)$$

Integration of Eq. 87 with the boundary condition that $\mathcal{E} = 0$ at $x = 0$ (assuming $\delta \rightarrow 0$) gives

$$\mathcal{E}^2 = \frac{2J}{\epsilon_s \mu} x. \quad (88)$$

Then

$$|\mathcal{E}| = \frac{dV}{dx} \sqrt{\frac{2Jx}{\epsilon_s \mu}} \quad (89)$$

or

$$dV = \sqrt{\frac{2J}{\epsilon_s \mu}} \sqrt{x} dx. \quad (90)$$

Integration of Eq. 90 with the boundary condition that $V = V$ at $x = L$ gives

$$V = \frac{2}{3} \left(\frac{2J}{\epsilon_s \mu} \right)^{1/2} L^{3/2}. \quad (91)$$

From Eq. 91 we obtain

$$J = \frac{9\epsilon_s \mu V^2}{8L^3} \sim V^2. \quad (92)$$

Therefore, the space-charge-limited current in the constant-mobility regime is proportional to the square of the applied voltage (Fig. 21b).

In the velocity-saturation regime, Eq. 83 becomes $J = qn v_s$, where v_s is the saturation velocity. Substituting $qn = J/v_s$ into Eq. 84 and using the same boundary conditions, we obtain the space-charge-limited current

$$J = \frac{2\epsilon_s v_s}{L^2} V \sim V. \quad (93)$$

Thus, in the velocity-saturation regime, the current varies linearly with the applied voltage.

► 2.8 HIGH-FIELD EFFECTS

At low electric-fields, the drift velocity is linearly proportional to the applied field. We assume that the time interval between collisions, τ_c , is independent of the applied field. This is a reasonable assumption as long as the drift velocity is small compared with the thermal velocity of carriers, which is about 10^7 cm/s for silicon at room temperature.

As the drift velocity approaches the thermal velocity, its field dependence on the electric field will begin to depart from the linear relationship given in Section 2.1. Figure 22 shows the measured drift velocities of electrons and holes in silicon as a function of the electric field. It is apparent that initially the field dependence of the drift velocity is linear, corresponding to a constant mobility. As the electric field is increased, the drift velocity increases less rapidly. At sufficiently large fields, the drift velocity approaches a saturation velocity. The experimental results can be approximated by the empirical expression⁸

$$v_n, v_p = \frac{v_s}{\left[1 + (\mathcal{E}_0 / \mathcal{E})^\gamma\right]^{1/\gamma}}, \quad (94)$$

where v_s is the saturation velocity (10^7 cm/s for Si at 300 K), \mathcal{E}_0 is a constant, equal to 7×10^3 V/cm for electrons and 2×10^4 V/cm for holes in high-purity silicon materials, and γ is 2 for electrons and 1 for holes. Velocity saturation at high fields is particularly likely for field-effect transistors (FETs) with very short channels. Even moderate voltage can result in a high field along the channel. This effect is discussed in Chapter 5 and 6.

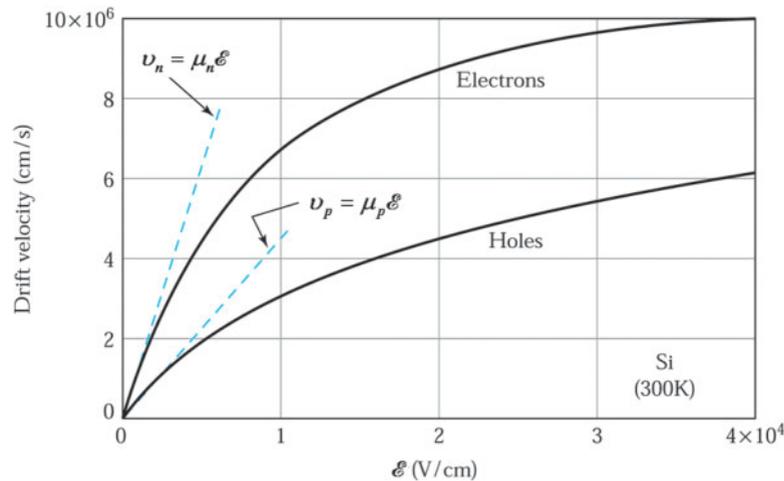


Fig. 22 Drift velocity versus electric field in Si.⁸

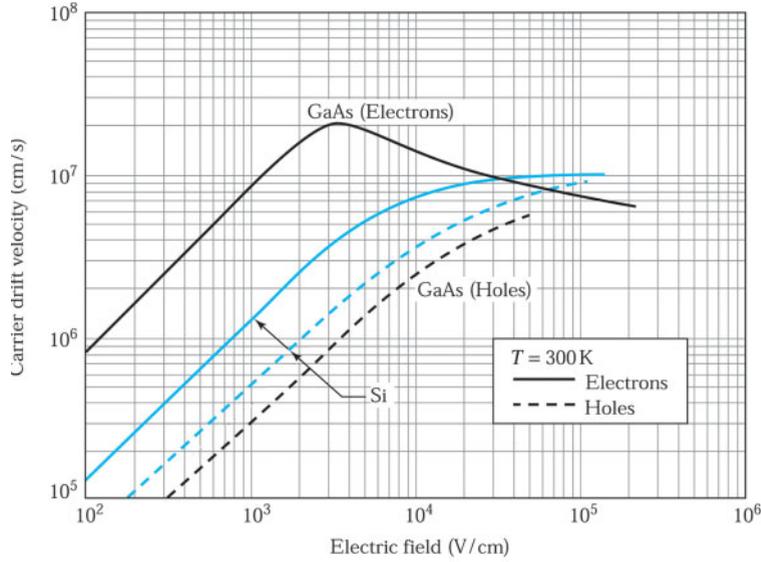


Fig. 23 Drift velocity versus electric field in Si and GaAs. Note that for *n*-type GaAs, there is a region of negative differential mobility.^{8,9}

The high-field transport in *n*-type gallium arsenide is different from that of silicon.⁹ Figure 23 shows the measured drift velocity versus field for *n*-type and *p*-type gallium arsenide. The results for silicon are also shown in this log-log plot for comparison. Note that for *n*-type GaAs, the drift velocity reaches a maximum, then decreases as the field further increases. This phenomenon is due to the energy-band structure of gallium arsenide, which allows the transfer of conduction electrons from a high-mobility energy minimum (called a valley) to low-mobility, higher-energy satellite valleys, that is, electron transfer from the central valley to the satellite valleys along the [111] direction shown in Fig. 14 of Chapter 1. This is called the transferred-electron effect.

To understand this phenomenon, consider the simple two-valley model of *n*-type gallium arsenide shown in Fig. 24. The energy separation between the two valleys is $\Delta E = 0.31$ eV. The lower valley's electron effective mass is denoted by m_1 , the electron mobility by μ_1 , and the electron density by n_1 . The upper-valley quantities are denoted by m_2 , μ_2 , and n_2 , respectively, and the total electron concentration is given by $n = n_1 + n_2$. The steady-state conductivity of the *n*-type GaAs can be written as

$$\sigma = q(\mu_1 n_1 + \mu_2 n_2) = qn\bar{\mu}, \quad (95)$$

where the average mobility is

$$\bar{\mu} \equiv (\mu_1 n_1 + \mu_2 n_2) / (n_1 + n_2). \quad (96)$$

The drift velocity is then

$$v_n = \bar{\mu} \mathcal{E}. \quad (97)$$

For simplicity we make the following assignments for the electron concentrations in the various ranges of electric-field values illustrated in Fig. 24. In Fig. 24*a*, the field is low and all electrons remain in the lower valley. In Fig. 24*b*, the field is higher and some electrons gain sufficient energies from the field to move to the higher valley. In Fig. 24*c*, the field is high enough to transfer all electrons to the higher valley. Thus, we have

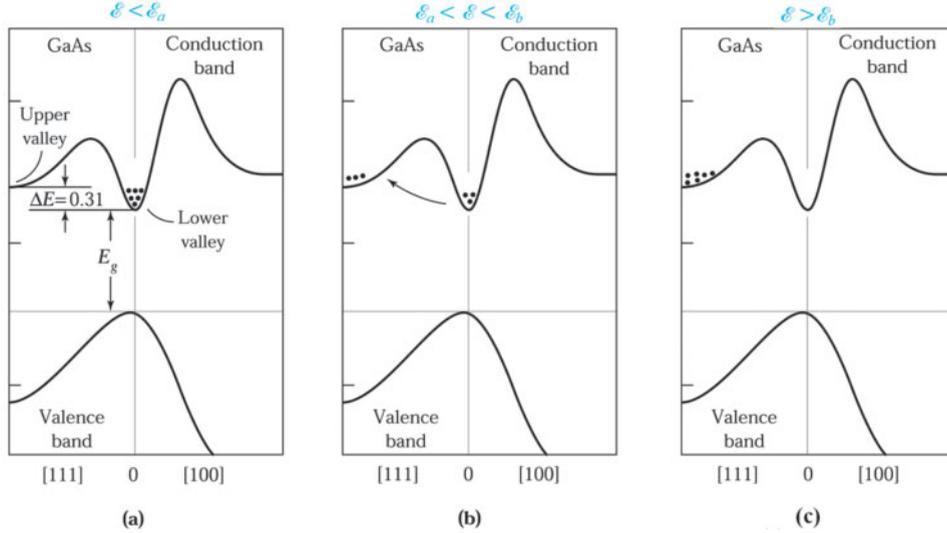


Fig. 24 Electron distributions under various conditions of electric fields: (a) $\mathcal{E} < \mathcal{E}_a$, (b) $\mathcal{E}_a < \mathcal{E} < \mathcal{E}_b$, and (c) $\mathcal{E} > \mathcal{E}_b$ for a two-valley semiconductor.

$$\begin{aligned}
 n_1 &\cong n & \text{and} & \quad n_2 \cong 0 & \text{for} & \quad 0 < \mathcal{E} < \mathcal{E}_a, \\
 n_1 + n_2 &\cong n & & & \text{for} & \quad \mathcal{E}_a < \mathcal{E} < \mathcal{E}_b, \\
 n_1 &\cong 0 & \text{and} & \quad n_2 \cong n & \text{for} & \quad \mathcal{E} > \mathcal{E}_b.
 \end{aligned} \tag{98}$$

Using these relations, the effective drift velocity takes on the asymptotic values

$$\begin{aligned}
 v_n &\cong \mu_1 \mathcal{E} & \text{for} & \quad 0 < \mathcal{E} < \mathcal{E}_a, \\
 v_n &\cong \mu_2 \mathcal{E} & \text{for} & \quad \mathcal{E} > \mathcal{E}_b.
 \end{aligned} \tag{99}$$

If $\mu_1 \mathcal{E}_a$ is larger than $\mu_2 \mathcal{E}_b$, there is a region in which the drift velocity decreases with increasing field between \mathcal{E}_a and \mathcal{E}_b , as shown in Fig. 25. Because of the characteristics of the drift velocity in n -type gallium arsenide, this material is used in the microwave transferred-electron devices discussed in Chapter 8.

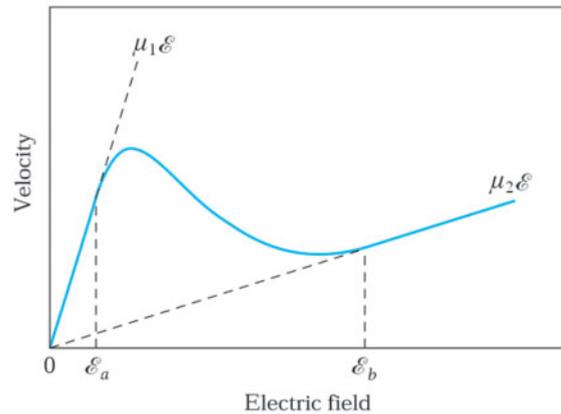


Fig. 25 One possible velocity-field characteristic of a two-valley semiconductor.

When the electric field in a semiconductor is increased above a certain value, the carriers gain enough kinetic energy to generate electron-hole pairs by the *avalanche process* shown schematically in Fig. 26. Consider an electron in the conduction band (designated by 1). If the electric field is high enough, this electron can gain kinetic energy before it collides with a valence electron. The high-energy electron in the conduction band can transfer some of its kinetic energy to the valence electron to make an upward transition to the conduction band. An electron-hole pair is generated (designated by 2 and 2'). Similarly, the generated pair begins to accelerate in the field and collides with other valence electrons, as indicated in the figure. In turn, they will generate other electron-hole pairs (e.g., 3 and 3', 4 and 4'), and so on. This process is called the avalanche process; it is also referred to as the *impact ionization process*. This process will result in breakdown in the *p-n* junction as discussed in Chapter 3.

To gain some ideas about the ionization energy involved, let us consider the process leading to 2–2' shown in Fig. 26. Just prior to the collision, the fast-moving electron (no. 1) has a kinetic energy $\frac{1}{2}m_1v_s^2$ and a momentum m_1v_s , where m_1 is the effective mass and v_s is the electron velocity. After collision, there are three carriers: the original electron plus an electron-hole pair (no. 2 and no. 2'). If we assume that the three carriers have the same effective mass, the same kinetic energy, and the same momentum, the total kinetic energy is $\frac{3}{2}m_1v_f^2$, and the total momentum is $3m_1v_f$, where v_f is the velocity after collision. To conserve both energy and momentum before and after the collision, we require that

$$\frac{1}{2}m_1v_s^2 = E_g + \frac{3}{2}m_1v_f^2 \tag{100}$$

and

$$m_1v_s = 3m_1v_f, \tag{101}$$

where in Eq. 100 the energy E_g is the bandgap corresponding to the minimum energy required to generate an electron-hole pair. Substituting Eq. 101 into Eq. 100 yields the required kinetic energy for the ionization process:

$$E_0 = \frac{1}{2}m_1v_s^2 = 1.5E_g. \tag{102}$$

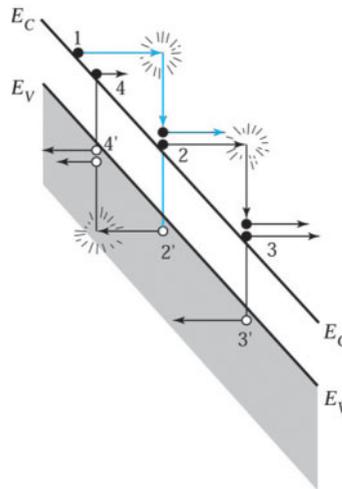


Fig. 26 Energy band diagram for the avalanche process.

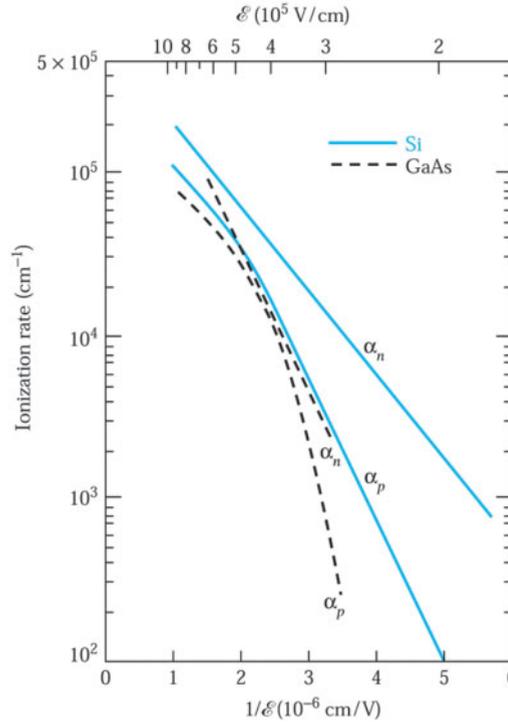


Fig. 27 Measured ionization rates versus reciprocal field for Si and GaAs.⁹

It is obvious that E_0 must be larger than the bandgap for the ionization process to occur. The actual energy required depends on the band structure. For silicon, the value for E_0 is 1.6 eV ($1.5 E_g$) for electrons and ~ 2.0 eV ($1.8 E_g$) for holes.

The number of electron-hole pairs generated by an electron per unit distance traveled is called the *ionization rate* for the electron, α_n . Similarly, α_p is the ionization rate for the holes. The measured ionization rates for silicon and gallium arsenide are shown⁹ in Fig. 27. We note that both α_n and α_p are strongly dependent on the electric field. For a substantially large ionization rate (say 10^4 cm^{-1}), the corresponding electric field is $\geq 3 \times 10^5 \text{ V/cm}$ for silicon and $\geq 4 \times 10^5 \text{ V/cm}$ for gallium arsenide. The electron-hole pair generation rate G_A from the avalanche process is given by

$$G_A = \frac{1}{q} (\alpha_n |J_n| + \alpha_p |J_p|), \quad (103)$$

where J_n and J_p are the electron and hole current densities, respectively. This expression can be used in the continuity equation for devices operated under an avalanche condition.

► SUMMARY

Various transport processes are at work in semiconductor devices. These include drift, diffusion, generation, recombination, thermionic emission, tunneling, space-charge effect, and impact ionization.

One of the key transport processes is the carrier drift under the influence of an electric field. At low fields, the drift velocity is proportional to the electric field. This proportionality constant is called mobility. Another key transport process is the carrier diffusion under the influence of the carrier concentration gradient. The total current is the sum of the drift and diffusion components.

Excess carriers in a semiconductor cause a nonequilibrium condition. Most semiconductor devices operate under nonequilibrium conditions. Carriers can be generated by various means such as forward biasing a $p-n$ junction, incident light, and impact ionization. The mechanism that restores equilibrium is the recombination of the excess minority carriers with the majority carriers by direct band-to-band recombination or via localized energy states in the forbidden energy gap. The governing equation for the rate of change of charge carriers is the continuity equation.

Among other transport processes, thermionic emission occurs when carriers in the surface region gain enough energy to be emitted into the vacuum level. Another, the tunneling process, is based on the quantum tunneling phenomena that results in the transport of electrons across a potential barrier even if the electron energy is smaller than the barrier height. In addition, we have the space-charge-limited current when the injected carriers are not compensated by the ionized impurities in the semiconductor.

As the electric field becomes higher, the drift velocity in silicon departs from its linear relationship with the applied field and approaches a saturation velocity. This effect is particularly important in the short-channel silicon field-effect transistors discussed in Chapter 5 and 6. The drift velocity in n -type GaAs reaches a maximum and then decreases as the field further increases. This is due to the transferred-electron effect, and the material is used in microwave devices discussed in Chapter 8. When the field exceeds a certain value, the carriers gain enough kinetic energy to generate electron-hole pair through Coulombic interaction. This effect is particularly important in the study of $p-n$ junctions. The high field accelerates these new electron-hole pairs, which collide with the lattice to create more electron-hole pairs. As this process, called impact ionization or the avalanche process, continues, the $p-n$ junction breaks down and conducts a large current. The junction breakdown is discussed in Chapter 3.

► REFERENCES

1. R. A. Smith, *Semiconductors*, 2nd ed., Cambridge Univ. Press, London, 1979.
2. J. L. Moll, *Physics of Semiconductors*, McGraw-Hill, New York, 1964.
3. W. F. Beadle, J. C. C. Tsai, and R. D. Plummer, Eds., *Quick Reference Manual for Semiconductor Engineers*, Wiley, New York, 1985.
4. (a) R. N. Hall, "Electron-Hole Recombination in Germanium," *Phys. Rev.*, **87**, 387. (1952); (b) W. Shockley and W. T. Read, "Statistics of Recombination of Holes and Electrons," *Phys. Rev.*, **87**, 835 (1952).
5. M. Prutton, *Surface Physics*, 2nd ed., Clarendon, Oxford, 1983.
6. A. S. Grove, *Physics and Technology of Semiconductor Devices*, Wiley, New York, 1967.
7. J. R. Haynes and W. Shockley, "The Mobility and Life of Injected Holes and Electrons in Germanium," *Phys. Rev.*, **81**, 835 (1951).
8. D. M. Caughey and R. E. Thomas, "Carrier Mobilities in Silicon Empirically Related to Doping and Field," *Proc. IEEE*, **55**, 2192 (1967).
9. S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, 3rd Ed., Wiley Interscience, Hoboken, 2007.
10. T. S. Moss, Ed., *Handbook on Semiconductors*, Vol. 1-4, North-Holland, Amsterdam, 1980.

► PROBLEMS (* DENOTES DIFFICULT PROBLEMS)

FOR SECTION 2.1 CARRIER DRIFT

1. Find the resistivities of intrinsic Si and intrinsic GaAs at 300 K.
2. Assume that the mobility of electrons in silicon at $T = 300$ K is $\mu_n = 1300$ cm²/V-s. Also assume that the mobility is mainly limited by lattice scattering. Determine the electron mobility at (a) $T = 200$ K and (b) $T = 400$ K.
3. Two scattering mechanisms exist in a semiconductor. If only the first mechanism is present, the mobility will be 250 cm²/V-s. If only the second mechanism is present, the mobility will be 500 cm²/V-s. Determine the mobility when both scattering mechanisms exist at the same time.
4. Find the electron and hole concentrations, mobilities, and resistivities of silicon samples at 300 K, for each of the following impurity concentrations: (a) 5×10^{15} boron atoms/cm³; (b) 2×10^{16} boron atoms/cm³ and 1.5×10^{16} arsenic atoms/cm³; and (c) 5×10^{15} boron atoms/cm³, 10^{17} arsenic atoms/cm³, and 10^{17} gallium atoms/cm³.
- *5. Consider a compensated n -type silicon at $T = 300$ K, with a conductivity of $\sigma = 16$ (Ω -cm)⁻¹ and an acceptor doping concentration of 10^{17} cm⁻³. Determine the donor concentration and the electron mobility. (A compensated semiconductor is one that contains both donor and acceptor impurity atoms in the same region.)
6. For a semiconductor with a constant mobility ratio $b \equiv \mu_n/\mu_p > 1$ independent of impurity concentration, find the maximum resistivity ρ_m in terms of the intrinsic resistivity ρ_i and the mobility ratio.
7. A four-point probe (with probe spacing of 0.5 mm) is used to measure the resistivity of a p -type silicon sample. Find the resistivity of the sample if its diameter is 200 μ m and its thickness is 50 μ m. The contact current is 1 mA, and the measured voltage between the inner two probes is 10 mV.
8. Given a silicon sample of unknown doping, Hall measurement provides the following information: $W = 0.05$ cm, $A = 1.6 \times 10^{-3}$ cm² (refer to Fig. 8), $I = 2.5$ mA, and the magnetic field is 30 nT (1 T = 10^{-4} Wb/cm²). If a Hall voltage of +10 mV is measured, find the Hall coefficient, conductivity type, majority carrier concentration, resistivity, and mobility of the semiconductor sample.
9. A semiconductor is doped with N_D ($N_D \gg n_i$) and has a resistance R_1 . The same semiconductor is then doped with an unknown amount of acceptors N_A ($N_A \gg N_D$), yielding a resistance of $0.5 R_1$. Find N_A in terms of N_D if $D/D_p = 50$.
- *10. Consider a semiconductor that is nonuniformly doped with donor impurity atoms $N_D(x)$.

Show that the induced electric field in the semiconductor in thermal equilibrium is given by

$$\mathcal{E}(x) = - \left(\frac{kT}{q} \right) \frac{1}{N_D(x)} \frac{dN_D(x)}{dx}.$$

FOR SECTION 2.2 CARRIER DIFFUSION

11. An intrinsic Si sample is doped with donors from one side such that $N_D = N_0 \exp(-ax)$. (a) Find an expression for the built-in field $\mathcal{E}(x)$ at equilibrium over the range for which $N_D \gg n_i$. (b) Evaluate $\mathcal{E}(x)$ when $a = 1$ μ m⁻¹.

12. An n -type Si slice of a thickness L is inhomogeneously doped with phosphorus donor whose concentration profile is given by $N_D(x) = N_0 + (N_L - N_0)(x/L)$. What is the formula for the electric potential difference between the front and the back surfaces when the sample is at thermal and electric equilibria regardless of how the mobility and diffusivity varies with position? What is the formula for the equilibrium electric field at a plane x from the front surface for a constant diffusivity and mobility?

FOR SECTION 2.3 GENERATION AND RECOMBINATION PROCESS

13. Calculate the electron and hole concentration under steady-state illumination in an n -type silicon with $G_L = 10^{16} \text{ cm}^{-3}\text{s}^{-1}$, $N_D = 10^{15} \text{ cm}^{-3}$, and $\tau_n = \tau_p = 10 \text{ }\mu\text{s}$.
14. An n -type silicon sample has 2×10^{16} arsenic atoms/cm³, 2×10^{15} bulk recombination centers/cm³, and 10^{10} surface recombination centers/cm². (a) Find the bulk minority carrier lifetime, the diffusion length, and the surface recombination velocity under low-injection conditions. The values of σ_p and σ_s are 5×10^{-15} and $2 \times 10^{-16} \text{ cm}^2$, respectively. (b) If the sample is illuminated with uniformly absorbed light that creates 10^{17} electron-hole pairs/cm²-s, what is the hole concentration at the surface?
15. Assume that an n -type semiconductor is uniformly illuminated, producing a uniform excess generation rate G . Show that in steady state the change in the semiconductor conductivity is given by $\Delta\sigma = q(\mu_n + \mu_p)\tau_p G$.

FOR SECTION 2.4 CONTINUITY EQUATION

16. The total current in a semiconductor is constant and is composed of electron drift current and hole diffusion current. The electron concentration is constant and equal to 10^{16} cm^{-3} . The hole concentration is given by

$$p(x) = 10^{15} \exp\left(\frac{-x}{L}\right) \text{ cm}^{-3} \quad (x \geq 0),$$

where $L = 12 \text{ }\mu\text{m}$. The hole diffusion coefficient is $D_p = 12 \text{ cm}^2/\text{s}$ and the electron mobility is $\mu_n = 1000 \text{ cm}^2/\text{V}\cdot\text{s}$. The total current density is $J = 4.8 \text{ A/cm}^2$. Calculate (a) the hole diffusion current density versus x , (b) the electron current density versus x , and (c) the electric field versus x .

- *17. Excess carriers are injected on one surface of a thin slice of n -type silicon with thickness W and extracted at the opposite surface where $p_n(W) = p_{no}$. There is no electric field in the region $0 < x < W$. Derive the expression for current densities at the two surfaces.
18. In Prob. 17, if carrier lifetime is $50 \text{ }\mu\text{s}$ and $W = 0.1 \text{ mm}$, calculate the portion of injected current that reaches the opposite surface by diffusion ($D = 50 \text{ cm}^2/\text{s}$).
- *19. An n -type semiconductor has excess carrier holes 10^{14} cm^{-3} , and a bulk minority carrier lifetime 10^{-6} s in the bulk material, and a minority carrier lifetime 10^{-7} s at the surface. Assume zero applied electric field and let $D_p = 10 \text{ cm}^2/\text{s}$. Determine the steady-state excess carrier concentration as a function of distance from the surface ($x = 0$) of the semiconductor.

FOR SECTION 2.5 THERMIONIC EMISSION PROCESS

20. A metal, with a work function $\phi_m = 4.2$ V, is deposited on an *n*-type silicon semiconductor with affinity $\chi = 4.0$ V and $E_g = 1.12$ eV. What is the potential barrier height seen by electrons in the metal moving into the semiconductor?
21. Consider a tungsten filament with metal work function ϕ_m inside a high vacuum chamber. Show that if a current is passed through the filament to heat it up sufficiently, the electrons with enough thermal energy will escape into the vacuum and the resulted thermionic current density is

$$J = A^* T^2 \exp\left(\frac{-q\phi_m}{kT}\right)$$

where A^* is $4\pi qmk^2 / h^3$ and m is free electron mass. The definite integral

$$\int_{-\infty}^{\infty} e^{-ax^2} dx = \left(\frac{\pi}{a}\right)^{1/2}.$$

FOR SECTION 2.6 TUNNELING PROCESS

22. Consider an electron with an energy of 2 eV impinging on a potential barrier with 20 eV and a width of 3 Å. What is the tunneling probability?
23. Evaluate the transmission coefficient for an electron of energy 2.2 eV impinging on a potential barrier of height 6.0 eV and thickness 10^{-10} meters. Repeat the calculation for a barrier thickness of 10^{-9} meters.

FOR SECTION 2.8 HIGH FIELD EFFECTS

24. Use the velocity-field relations for Si and GaAs shown in Fig. 23 to determine the transit time of electrons through a 1 μm distance in these materials for an electric field of (a) 1 kV/cm and (b) 50 kV/cm.
25. Assume that a conduction electron in Si ($\mu_n = 1350$ cm²/V-s) has a thermal energy kT , related to its mean thermal velocity by $E_{th} = m_0 v_{th}^2 / 2$. This electron is placed in an electric field of 100 V/cm. Show that the drift velocity of the electron in this case is small compared to its thermal velocity. Repeat for a field of 10^4 V/cm, using the same value of μ_n . Comment on the actual mobility effects at this higher value of field.

p - n Junction

- ▶ 3.1 THERMAL EQUILIBRIUM CONDITION
- ▶ 3.2 DEPLETION REGION
- ▶ 3.3 DEPLETION CAPACITANCE
- ▶ 3.4 CURRENT-VOLTAGE CHARACTERISTICS
- ▶ 3.5 CHARGE STORAGE AND TRANSIENT BEHAVIOR
- ▶ 3.6 JUNCTION BREAKDOWN
- ▶ 3.7 HETEROJUNCTION
- ▶ SUMMARY

In the preceding chapters we considered the carrier concentrations and transport phenomena in homogeneous semiconductor materials. In this chapter we discuss the behavior of single-crystal semiconductor material containing both p - and n -type regions that form a p - n junction.

A p - n junction serves an important role both in modern electronic applications and in understanding other semiconductor devices. It is used extensively in rectification, switching, and other operations in electronic circuits. It is a key building block for the bipolar transistor and thyristor (Chapter 4), as well as for metal-oxide-semiconductor field-effect transistors (MOSFETs) (Chapters 5 and 6). Given proper biasing conditions or when exposed to light, the p - n junction also functions as either a microwave (Chapter 8) or photonic device (Chapters 9 and 10).

We also consider a related device—the heterojunction, which is a junction formed between two dissimilar semiconductors. It has many unique features that are not readily available from the conventional p - n junction. The heterojunction is an important building block for heterojunction bipolar transistors (Chapter 4), modulation doped field-effect transistors (Chapter 7), quantum-effect devices (Chapter 8), and photonic devices (Chapters 9 and 10).

Specifically, we cover the following topics:

- The band diagram of a p - n junction at thermal equilibrium.
- The behavior of the junction depletion layer under voltage biases.
- The current transport in a p - n junction and the influence of the generation and recombination processes.
- The charge storage in a p - n junction and its influence on the transient behavior.
- The avalanche multiplication in a p - n junction and its impact on the maximum reverse voltage
- The heterojunction and its basic characteristics.

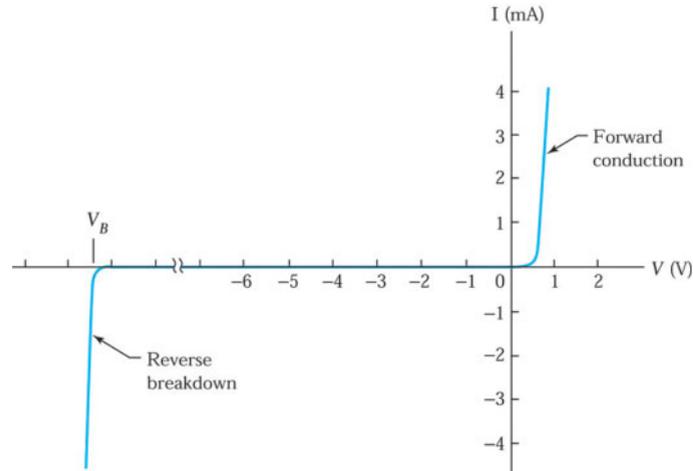


Fig. 1 Current-voltage characteristics of a typical silicon *p-n* junction.

► 3.1 THERMAL EQUILIBRIUM CONDITION

Today, planar technology is used extensively for *p-n* junction and integrated circuit (IC) fabrication. The planar processes include oxidation, lithography, ion implantation, and metallization; they will be discussed in Chapters 11-15. The most important characteristic of *p-n* junctions is that they rectify: that is, they allow current to flow easily in only one direction. Figure 1 shows the current-voltage characteristics of a typical silicon *p-n* junction. When we apply “forward bias” to the junction (i.e., positive voltage on the *p*-side), the current increases rapidly as the voltage increases. However, when we apply a “reverse bias,” virtually no current flows initially. As the reverse bias is increased the current remains very small until a critical voltage is reached, at which point the current suddenly increases. This sudden increase in current is referred to as the junction breakdown. The applied forward voltage is usually less than 1V, but the reverse critical voltage, or breakdown voltage, can vary from just a few volts to many thousands of volts, depending on the doping concentration and other device parameters.

1.1.1 Band Diagram

In Fig. 2a, we see two regions of *p*- and *n*-type semiconductor materials that are uniformly doped and physically separated before the junction is formed. Note that the Fermi level E_F is near the valence band edge in the *p*-type material and near the conduction band edge in the *n*-type material. While *p*-type material contains a large concentration of holes with few electrons, the opposite is true for *n*-type material.

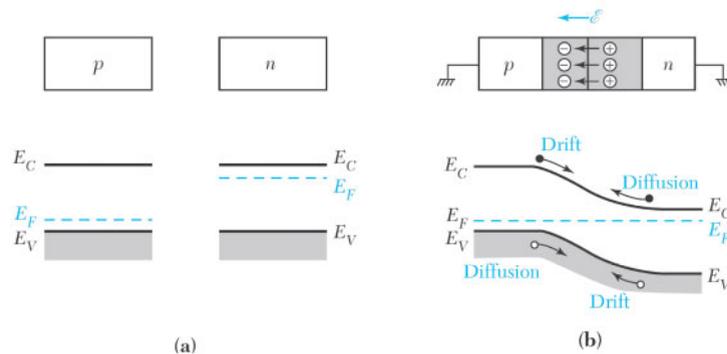


Fig. 2 (a) Uniformly doped *p*-type and *n*-type semiconductors before the junction is formed. (b) The electric field in the depletion region and the energy band diagram of a *p-n* junction in thermal equilibrium.

When the p - and n -type semiconductors are jointed together, the large carrier concentration gradients at the junction cause carrier diffusion. Holes from the p -side diffuse into the n -side, and electrons from the n -side diffuse into the p -side. As holes continue to leave the p -side, some of the negative acceptor ions (N_A^-) near the junction are left uncompensated because the acceptors are fixed in the semiconductor lattice, whereas the holes are mobile. Similarly, some of the positive donor ions (N_D^+) near the junction are left uncompensated as the electrons leave the n -side. Consequently, a negative space charge forms near the p -side of the junction and a positive space charge forms near the n -side. This space charge region creates an electric field that is directed from the positive charge toward the negative charge, as indicated in the upper illustration of Fig. 2*b*.

The electric field is in the direction opposite to the diffusion current for each type of charge carrier. The lower illustration of Fig. 2*b* shows that the hole diffusion current flows from left to right, whereas the hole drift current due to the electric field flows from right to left. The electron diffusion current also flows from left to right, whereas the electron drift current flows in the opposite direction. Note that because of their negative charge, electrons diffuse from right to left, opposite to the direction of electron current.

3.1.2 Equilibrium Fermi Levels

At thermal equilibrium, i.e. the steady-state condition at a given temperature with no external excitations, the individual electron and hole currents flowing across the junctions are identically zero. Thus, for each type of carrier the drift current due to the electric field must exactly cancel the diffusion current due to the concentration gradient. From Eq. 32 in Chapter 2,

$$\begin{aligned} J_p &= J_p(\text{drift}) + J_p(\text{diffusion}) \\ &= q\mu_p p \mathcal{E} - qD_p \frac{dp}{dx} \\ &= q\mu_p p \left(\frac{1}{q} \frac{dE_i}{dx} \right) - kT\mu_p \frac{dp}{dx} = 0, \end{aligned} \quad (1)$$

where we have used Eq. 8 of Chapter 2 for the electric field and the Einstein relation $D_p = (kT/q)\mu_p$. Substituting the expression for hole concentration

$$p = n_i e^{(E_i - E_F)/kT} \quad (2)$$

and its derivative

$$\frac{dp}{dx} = \frac{p}{kT} \left(\frac{dE_i}{dx} - \frac{dE_F}{dx} \right) \quad (3)$$

into Eq. 1 yields the net hole current density

$$J_p = \mu_p p \frac{dE_F}{dx} = 0 \quad (4)$$

or

$$\frac{dE_F}{dx} = 0. \quad (5)$$

Similarly, we obtain for the net electron current density

$$\begin{aligned} J_n &= J_n(\text{drift}) + J_n(\text{diffusion}) \\ &= q\mu_n n \mathcal{E} + qD_n \frac{dn}{dx} \\ &= \mu_n n \frac{dE_F}{dx} = 0. \end{aligned} \quad (6)$$

or

$$\frac{dE_F}{dx} = 0.$$

Thus, for the condition of zero net electron and hole currents, the Fermi level must be constant (i.e., independent of x) throughout the sample, as illustrated in the energy band diagram of Fig. 2*b*.

The constant Fermi level required at thermal equilibrium results in a unique space charge distribution at the junction. We repeat the one-dimensional *p-n* junction and the corresponding equilibrium energy band diagram in Figs. 3*a* and 3*b*, respectively. The unique space charge distribution and the electrostatic potential ψ are given by Poisson's equation:

$$\boxed{\frac{d^2\psi}{dx^2} \equiv \frac{d\mathcal{E}}{dx} = \frac{\rho_s}{\epsilon_s} = \frac{q}{\epsilon_s} (N_D - N_A + p - n).} \quad (7)$$

Here we assume that all donors and acceptors are ionized.

In regions far away from the metallurgical junction, charge neutrality is maintained and the total space charge density is zero. For these neutral regions we can simplify Eq. 7 to

$$\frac{d^2\psi}{dx^2} = 0 \quad (8)$$

and

$$N_D - N_A + p - n = 0. \quad (9)$$

For a *p*-type neutral region, we assume $N_D = 0$ and $p \gg n$. The electrostatic potential of the *p*-type neutral region with respect to the Fermi level, designated as ψ_p in Fig. 3*b*, can be obtained by setting $N_D = n = 0$ in Eq. 9 and substituting the result ($p = N_A$) into Eq. 2:

$$\psi_p \equiv -\frac{1}{q} (E_i - E_F) \Big|_{x \leq x_p} = -\frac{kT}{q} \ln \left(\frac{N_A}{n_i} \right). \quad (10)$$

Similarly, we obtain the electrostatic potential of the *n*-type neutral region with respect to the Fermi level:

$$\psi_n \equiv -\frac{1}{q} (E_i - E_F) \Big|_{x \geq x_n} = \frac{kT}{q} \ln \left(\frac{N_D}{n_i} \right). \quad (11)$$

The total electrostatic potential difference between the *p*-side and the *n*-side neutral regions at thermal equilibrium is called the *built-in potential* V_{bi} :

$$\boxed{V_{bi} = \psi_n - \psi_p = \frac{kT}{q} \ln \left(\frac{N_A N_D}{n_i^2} \right).} \quad (12)$$

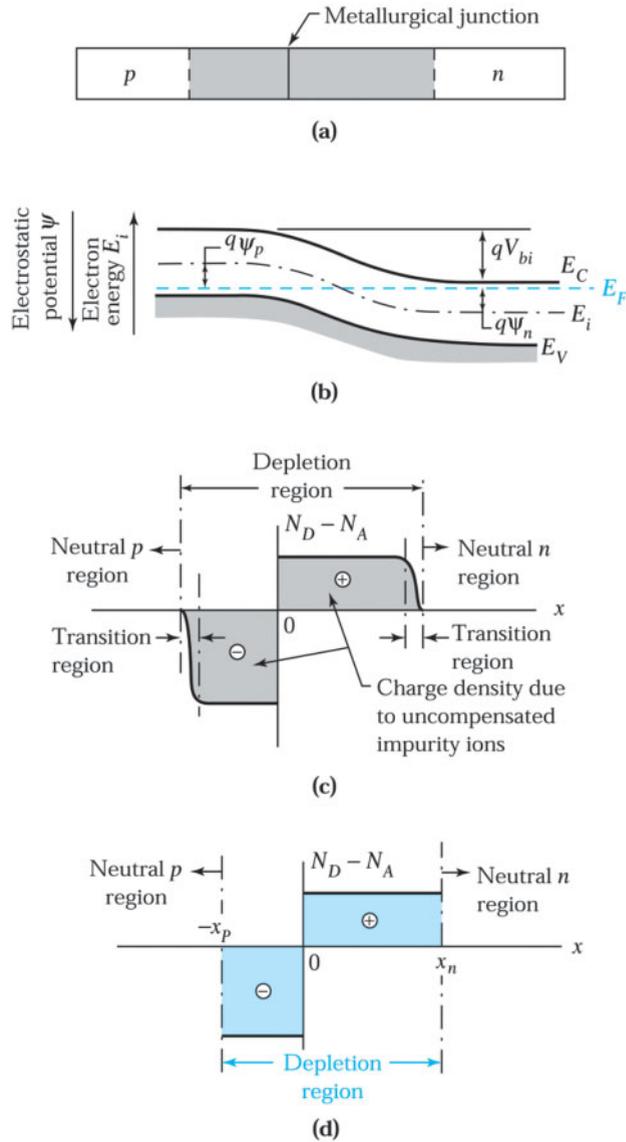


Fig. 3 (a) A p - n junction with abrupt doping changes at the metallurgical junction. (b) Energy band diagram of an abrupt junction at thermal equilibrium. (c) Space charge distribution. (d) Rectangular approximation of the space charge distribution.

3.1.3 Space Charge

Moving from a neutral region toward the junction, we encounter the narrow transition region shown in Fig. 3c. Here the space charge of impurity ions is partially compensated by the mobile carriers. Beyond the transition region we enter the completely depleted region where the mobile carrier densities are zero. This is called the *depletion region* (also the space-charge region). For typical p - n junctions in silicon and gallium arsenide, the width of each transition region is small compared with the width of the depletion region. Therefore, we can neglect the transition region and represent the depletion region by the rectangular distribution shown in Fig. 3d, where x_p and x_n denote the depletion layer widths of the p - and n -sides for the completely depleted region with $p = n = 0$. Equation 7 becomes

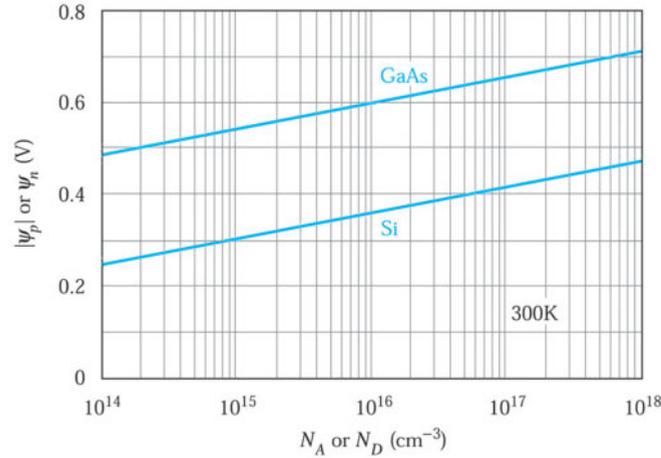


Fig. 4 Electrostatic potentials on the *p*-side and *n*-side of abrupt junctions in Si and GaAs as a function of impurity concentration.

$$\frac{d^2\psi}{dx^2} = \frac{q}{\epsilon_s}(N_A - N_D) \tag{13}$$

The magnitudes of $|\psi_p|$ and ψ_n as calculated from Eqs.10 and 11 are plotted in Fig. 4 as a function of the doping concentration of silicon and gallium arsenide. For a given doping concentration, the electrostatic potential of gallium arsenide is higher because of its smaller intrinsic concentration n_i .

► **EXAMPLE 1**

Calculate the built-in potential for a silicon *p*–*n* junction with $N_A = 10^{18} \text{ cm}^{-3}$ and $N_D = 10^{15} \text{ cm}^{-3}$ at 300 K.

SOLUTION From Eq. 12 we obtain

$$V_{bi} = (0.0259) \ln \left[\frac{10^{18} \times 10^{15}}{(9.65 \times 10^9)^2} \right] = 0.774 \text{ V.}$$

Also from Fig. 4,

$$V_{bi} = \psi_n + |\psi_p| = 0.30 \text{ V} + 0.47 \text{ V} = 0.77 \text{ V.} \quad \blacktriangleleft$$

► **3.2 DEPLETION REGION**

To solve Poisson’s equation (Eq. 13) we must know the impurity distribution. In this section we consider two important cases—the abrupt junction and the linearly graded junction. Figure 5a shows an *abrupt junction*, that is, a *p*–*n* junction formed by shallow diffusion or low-energy ion implantation. The impurity distribution of the junction can be approximated by an abrupt transition of doping concentration between the *n*- and *p*-type regions. Figure 5b shows a linearly graded junction. For either deep diffusions or high-energy ion implantations, the impurity profiles may be approximated by linearly graded junctions: that is, the impurity distribution varies linearly across the junction. We consider the depletion regions of both types of junction.

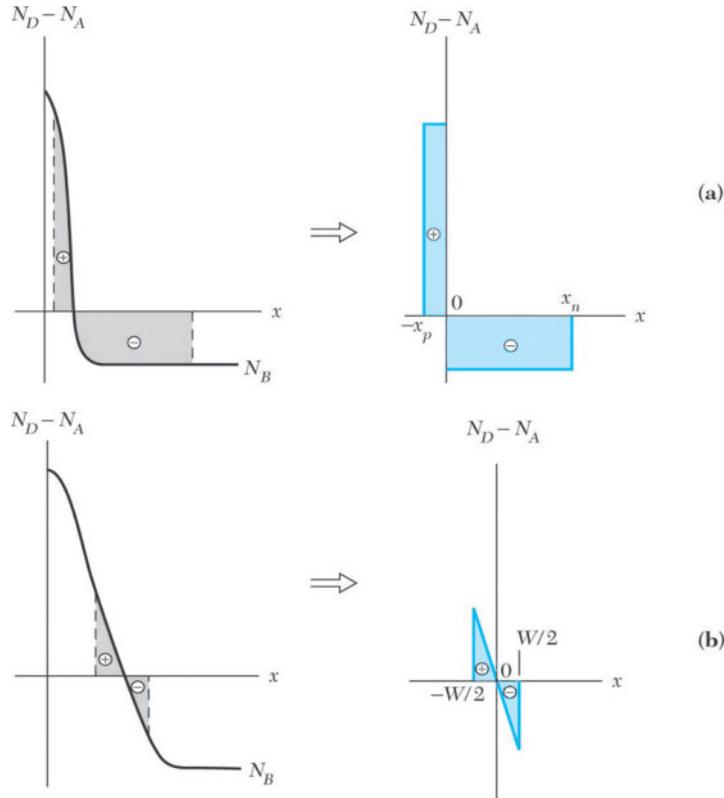


Fig. 5 Approximate doping profiles. (a) Abrupt junction. (b) Linearly graded junction.

3.2.1 Abrupt Junction

The space charge distribution of an abrupt junction is shown in Fig. 6a. In the depletion region, free carriers are totally depleted so that Poisson's equation (Eq. 13) simplifies to

$$\frac{d^2\psi}{dx^2} + \frac{qN_A}{\epsilon_s} \quad \text{for } -x_p \leq x < 0, \quad (14a)$$

$$\frac{d^2\psi}{dx^2} - \frac{qN_D}{\epsilon_s} \quad \text{for } 0 < x \leq x_n. \quad (14b)$$

The overall space charge neutrality of the semiconductor requires that the total negative space charge per unit area in the p -side must precisely equal the total positive space charge per unit area in the n -side:

$$N_A x_p = N_D x_n. \quad (15)$$

The total depletion layer width W is given by

$$W = x_p + x_n. \quad (16)$$

The electric field shown in Fig. 6b is obtained by integrating Eqs. 14a and 14b, which gives

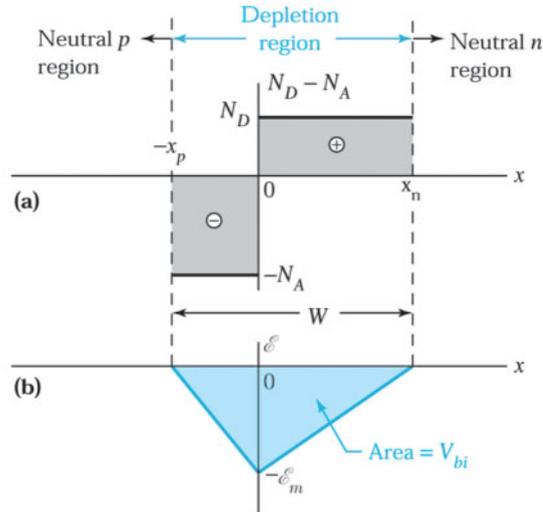


Fig. 6 (a) Space charge distribution in the depletion region at thermal equilibrium. (b) Electric-field distribution. The shaded area corresponds to the built-in potential.

$$\mathcal{E}(x) = \frac{d\psi}{dx} = \frac{qN_A(x + x_p)}{\epsilon_s} \quad \text{for } x_p \leq x < 0 \quad (17a)$$

and

$$\mathcal{E}(x) = \mathcal{E}_m + \frac{qN_D x}{\epsilon_s} = \frac{qN_D}{\epsilon_s}(x - x_n) \quad \text{for } 0 < x \leq x_n, \quad (17b)$$

where \mathcal{E}_m is the maximum field that exists at $x = 0$ and is given by

$$\mathcal{E}_m = \frac{qN_D x_n}{\epsilon_s} = \frac{qN_A x_p}{\epsilon_s}. \quad (18)$$

Integrating Eqs. 17a and 17b over the depletion region gives the total potential variation, namely, the built-in potential V_{bi} :

$$\begin{aligned} V_{bi} &= -\int_{x_p}^{x_n} \mathcal{E}(x) dx = -\int_{x_p}^0 \mathcal{E}(x) dx \Big|_{p \text{ side}} - \int_0^{x_n} \mathcal{E}(x) dx \Big|_{n \text{ side}} \\ &= \frac{qN_A x_p^2}{2\epsilon_s} + \frac{qN_D x_n^2}{2\epsilon_s} = \frac{1}{2} \mathcal{E}_m W. \end{aligned} \quad (19)$$

Therefore, the area of the field triangle in Fig. 6b corresponds to the built-in potential.

Combining Eqs. 15 through 19 gives the total depletion layer width as a function of the built-in potential,

$$W = \sqrt{\frac{2\epsilon_s}{q} \left(\frac{N_A + N_D}{N_A N_D} \right) V_{bi}}. \quad (20)$$

When the impurity concentration on one side of an abrupt junction is much higher than that on the other side, the junction is called a *one-sided abrupt junction* (Fig. 7a). Figure 7b shows the space charge distribution of a one-sided abrupt p^+-n junction, where $N_A \gg N_D$. In this case, the depletion layer width of the p -side is much smaller than that of the n -side (i.e., $x_p \ll x_n$), and the expression for W can be simplified to

$$W \cong x_n = \sqrt{\frac{2\epsilon_s V_{bi}}{qN_D}}. \quad (21)$$

The expression for the electric-field distribution is the same as Eq. 17b:

$$\mathcal{E}(x) = -\mathcal{E}_m + \frac{qN_B x}{\epsilon_s}, \quad (22)$$

where N_B is the lightly doped bulk concentration (i.e., N_D for a p^+-n junction). The field decreases to zero at $x = W$. Therefore,

$$\mathcal{E}_m = \frac{qN_B W}{\epsilon_s} \quad (23)$$

and

$$\mathcal{E}(x) = \frac{qN_B}{\epsilon_s} (-W + x) = -\mathcal{E}_m \left(1 - \frac{x}{W}\right), \quad (24)$$

as shown in Fig. 7c.

Integrating Poisson's equation once more gives the potential distribution

$$\psi(x) = \int_0^x \mathcal{E} dx = \mathcal{E}_m \left(x - \frac{x^2}{2W}\right) + \text{constant}. \quad (25)$$

With zero potential in the neutral p -region as a reference, or $\psi(0) = 0$, and employing Eq. 19, we have

$$\psi(x) = \frac{V_{bi} x}{W} \left(2 - \frac{x}{W}\right). \quad (26)$$

The potential distribution is shown in Fig. 7d.

► EXAMPLE 2

For a silicon one-sided abrupt junction with $N_A = 10^{19} \text{ cm}^{-3}$ and $N_D = 10^{16} \text{ cm}^{-3}$, calculate the depletion layer width and the maximum field at zero bias ($T = 300 \text{ K}$).

SOLUTION From Eqs. 12, 21, and 23, we obtain

$$\begin{aligned} V_{bi} &= 0.0259 \ln \left[\frac{10^{19} \times 10^{16}}{(9.65 \times 10^9)^2} \right] = 0.895 \text{ V}, \\ W &\cong \sqrt{\frac{2\epsilon_s V_{bi}}{qN_D}} = 3.41 \times 10^{-5} \text{ cm} = 0.341 \text{ } \mu\text{m} \\ \mathcal{E}_m &= \frac{qN_B W}{\epsilon_s} = 0.52 \times 10^4 \text{ V/cm}. \end{aligned}$$

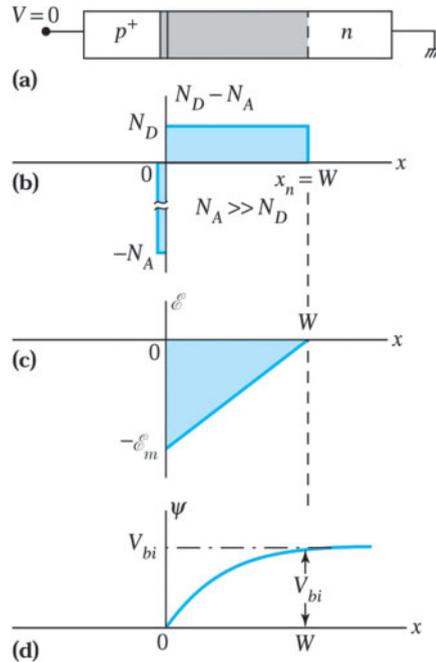


Fig. 7 (a) One-sided abrupt junction (with $N_A \gg N_D$) in thermal equilibrium. (b) Space charge distribution. (c) Electric-field distribution. (d) Potential distribution with distance, where V_{bi} is the built-in potential.

The previous discussions are for a *p-n* junction at thermal equilibrium without external bias. The equilibrium energy band diagram, shown again in Fig. 8a, illustrates that the total electrostatic potential across the junction is V_{bi} . The corresponding potential energy difference from the *p*-side to the *n*-side is qV_{bi} . If we apply a positive voltage V_F to the *p*-side with respect to the *n*-side, the *p-n* junction becomes forward-biased, as shown in Fig. 8b. The total electrostatic potential across the junction decreases by V_F ; that is, it is replaced with $V_{bi} - V_F$. Thus, forward bias reduces the depletion layer width.

By contrast, as shown in Fig. 8c, if we apply positive voltage V_R to the *n*-side with respect to the *p*-side, the *p-n* junction now becomes reverse-biased and the total electrostatic potential across the junction increases by V_R ; that is, it is replaced by $V_{bi} + V_R$. Here, we find that reverse bias increases the depletion layer width. Substituting these voltage values in Eq. 21 yields the depletion layer widths as a function of the applied voltage for a one-sided abrupt junction:

$$W = \sqrt{\frac{2\epsilon_s (V_{bi} - V)}{qN_B}}, \quad (27)$$

where N_B is the lightly doped bulk concentration and V is positive for forward bias and negative for reverse bias. Note that the depletion layer width W varies as the square root of the total electrostatic potential difference across the junction.

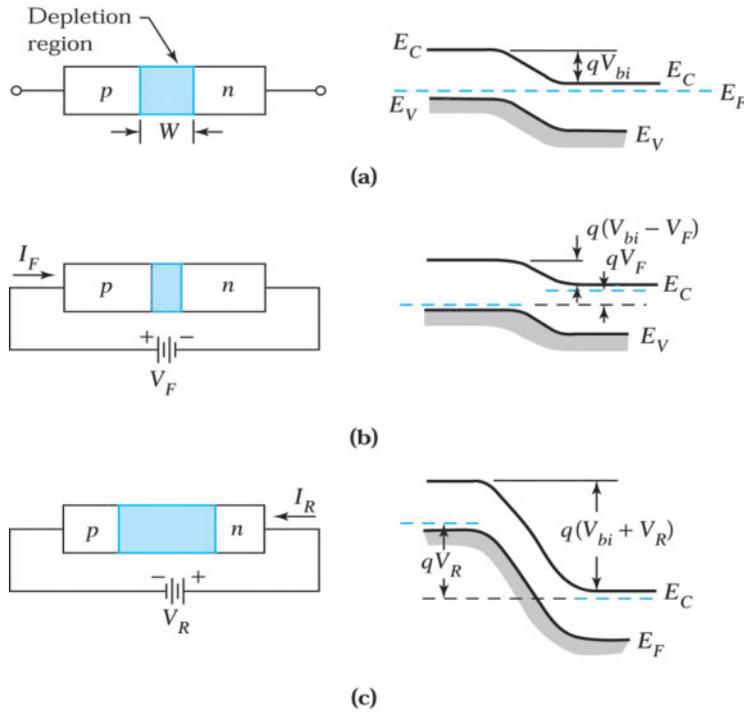


Fig. 8 Schematic representations of depletion layer width and energy band diagrams of a p - n junction under various biasing conditions. (a) Thermal-equilibrium condition. (b) Forward-bias condition. (c) Reverse-bias condition.

3.2.2 Linearly Graded Junction

We first consider the case of thermal equilibrium. The impurity distribution for a linearly graded junction is shown in Fig. 9a. The Poisson's equation for the case is

$$\frac{d^2\psi}{dx^2} = \frac{d\mathcal{E}}{dx} = \frac{\rho_s}{\epsilon_s} = \frac{q}{\epsilon_s} ax \quad -\frac{W}{2} \leq x \leq \frac{W}{2}, \quad (28)$$

where a is the impurity gradient (in cm^{-4}) and W is the depletion-layer width.

We have assumed that mobile carriers are negligible in the depletion region. By integrating Eq. 28 once with the boundary conditions that the electric field is zero at $\pm W/2$, we obtain the electric-field distribution shown in Fig. 9b:

$$\mathcal{E}(x) = -\frac{qa}{\epsilon_s} \left[\frac{(W/2)^2 - x^2}{2} \right]. \quad (29)$$

The maximum field at $x = 0$ is

$$\mathcal{E}_m = \frac{qaW^2}{8\epsilon_s}. \quad (29a)$$

Integrating Eq. 28 once again yields both the potential distribution and the corresponding energy band diagram shown in Figs. 9c and 9d, respectively. The built-in potential and the depletion layer width are given by

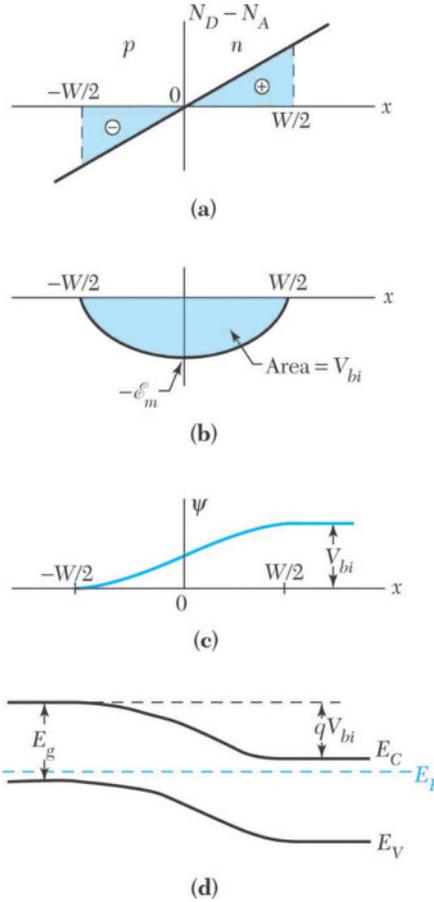


Fig. 9 Linearly graded junction in thermal equilibrium. (a) Impurity distribution. (b) Electric field distribution. (c) Potential distribution. (d) Energy band diagram.

$$V_{bi} = \frac{qaW^3}{12\epsilon_s} \tag{30}$$

and

$$W = \left(\frac{12\epsilon_s V_{bi}}{qa} \right)^{1/3} . \tag{31}$$

Since the values of the impurity concentrations at the edges of the depletion region ($-W/2$ and $W/2$) are the same and both are equal to $aW/2$, the built-in potential for a linearly graded junction may be expressed in a form similar to Eq. 12*:

$$V_{bi} = \frac{kT}{q} \ln \left[\frac{(aW/2)(aW/2)}{n_i^2} \right] = \frac{2kT}{q} \ln \left(\frac{aW}{2n_i} \right) . \tag{32}$$

* Based on an accurate numerical technique, the built-in potential is given by $V_{bi} = \frac{2kT}{3q} \ln \left(\frac{a^2 \epsilon_s kT / q}{8qn_i^3} \right)$. For a given impurity gradient, V_{bi} is smaller than that calculated from Eq.32 by about 0.05 – 0.1 V.

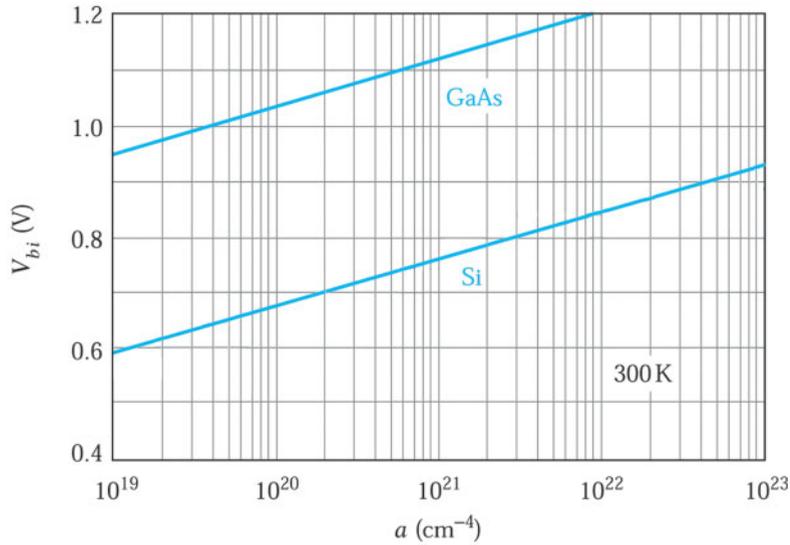


Fig. 10 Built-in potential for a linearly graded junction in Si and GaAs as a function of impurity gradient.

Solving the transcendental equation that results when W is eliminated from Eqs. 31 and 32 yields the built-in potential as a function of a . The results for silicon and gallium arsenide linearly graded junctions are shown in Fig. 10.

When either forward or reverse bias is applied to the linearly graded junction, the variations of the depletion layer width and the energy band diagram will be similar to those shown in Fig. 8 for abrupt junctions. However, the depletion layer width will vary as $(V_{bi} - V)^{1/3}$, where V is positive for forward bias and negative for reverse bias.

► EXAMPLE 3

For a silicon linearly graded junction with an impurity gradient of 10^{20} cm^{-4} , calculate the depletion-layer width and the maximum field and built-in voltage ($T = 300 \text{ K}$).

SOLUTION To solve the transcendental equation of Eq. 32, we can use a simple numeric method to find the built-in potential V_{bi} and the depletion layer width W simultaneously. We start with a reasonable V_{bi} value, e.g., 0.3 V, to calculate W by Eq. 31. W is 0.619 μm . Then, W and a are substituted in Eq. 32 to obtain V_{bi} . V_{bi} is 0.657 V. This V_{bi} value is substituted in Eq. 31 to obtain a new W , and in turn to calculate a new V_{bi} . After several trials as shown in the following table, we obtain the final V_{bi} of 0.671 V and W of 0.809 μm .

| trial | 1 | 2 | 3 | 4 | 5 |
|-----------------------|-------|-------|-------|-------|-------|
| V_{bi} (V) | 0.3 | 0.657 | 0.670 | 0.671 | 0.671 |
| W (μm) | 0.619 | 0.804 | 0.809 | 0.809 | 0.809 |

$$\mathcal{E}_m = \frac{qaW^2}{8\epsilon_s} = \frac{1.6 \times 10^{-19} \times 10^{20} \times (0.809 \times 10^{-4})^2}{8 \times 11.9 \times 8.85 \times 10^{-14}} = 1.24 \times 10^4 \text{ V/cm}$$

► 3.3 DEPLETION CAPACITANCE

The junction depletion-layer capacitance per unit area is defined as $C_j = dQ/dV$, where dQ is the incremental change in depletion-layer charge per unit area for an incremental change in the applied voltage dV .*

Figure 11 illustrates the depletion capacitance of a *p-n* junction with an arbitrary impurity distribution. The charge and electric-field distributions indicated by the solid lines correspond to a voltage V applied to the *n*-side. If this voltage is increased by an amount dV , the charge and field distributions will expand to those regions bounded by the dashed lines.

In Fig. 11*b*, the incremental charge dQ corresponds to the colored area between the two charge distribution curves on either side of the depletion region. The incremental space charges on the *n*- and *p*-sides of the depletion region are equal but with an opposite charge polarity, thus maintaining overall charge neutrality. This incremental charge dQ causes an increase in the electric field by an amount $d\mathcal{E} = dQ/\epsilon_s$ (from Poisson's equation). The corresponding change in the applied voltage dV , represented by the colored area in Fig. 11*c*, is approximately $Wd\mathcal{E}$, which equals WdQ/ϵ_s . Therefore, the depletion capacitance per unit area is given by

$$C_j \equiv \frac{dQ}{dV} = \frac{dQ}{W \frac{dQ}{\epsilon_s}} = \frac{\epsilon_s}{W} \tag{33}$$

or

$$C_j = \frac{\epsilon_s}{W} \text{ F/cm}^2. \tag{33a}$$

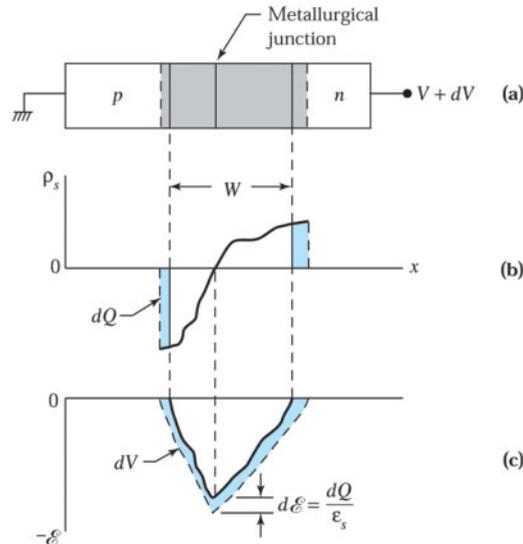


Fig. 11 (a) *p-n* junction with an arbitrary impurity profile under reverse bias. (b) Change in space charge distribution due to change in applied bias. (c) Corresponding change in electric-field distribution.

*The capacitance is also referred to as the transition region capacitance.

3.3.1 Capacitance-Voltage Characteristics

Equation 33 for the depletion capacitance per unit area is the same as the standard expression for a parallel-plate capacitor where the spacing between the two plates represents the depletion-layer width. The equation is valid for any arbitrary impurity distribution.

In deriving Eq. 33 we have assumed that only the variation of the space charge in the depletion region contributes to the capacitance. This certainly is a good assumption for the reverse-bias condition. For forward biases, however, a large current can flow across the junction, corresponding to a large number of mobile carriers present within the neutral region. The incremental change of these mobile carriers with respect to the biasing voltage contributes an additional term, called the diffusion capacitance, which is considered in Section 3.5.

For a one-sided abrupt junction, we obtain, from Eqs. 27 and 33,

$$C_j = \frac{\epsilon_s}{W} \sqrt{\frac{q\epsilon_s N_B}{2(V_{bi} - V)}} \quad (34)$$

or

$$\boxed{\frac{1}{C_j^2} = \frac{2(V_{bi} - V)}{q\epsilon_s N_B}} \quad (35)$$

It is clear from Eq. 35 that a plot of $1/C_j^2$ versus V produces a straight line for a one-sided abrupt junction. The slope gives the impurity concentration N_B of the substrate, and the intercept (at $1/C_j^2 = 0$) gives V_{bi} .

► EXAMPLE 4

For a silicon one-sided abrupt junction with $N_A = 2 \times 10^{19} \text{ cm}^{-3}$ and $N_D = 8 \times 10^{15} \text{ cm}^{-3}$, calculate the junction capacitance at zero bias and a reverse bias of 4 V ($T = 300 \text{ K}$).

SOLUTION From Eqs. 12, 27, and 34, we obtain at zero bias

$$V_{bi} = 0.0259 \ln \left[\frac{2 \times 10^{19} \times 8 \times 10^{15}}{(9.65 \times 10^9)^2} \right] = 0.906 \text{ V}$$

$$W|_{V=0} \cong \sqrt{\frac{2\epsilon_s V_{bi}}{qN_D}} = \sqrt{\frac{2 \times 11.9 \times 8.85 \times 10^{-14} \times 0.906}{1.6 \times 10^{-19} \times 8 \times 10^{15}}} = 3.86 \times 10^{-5} \text{ cm} = 0.386 \text{ }\mu\text{m}$$

$$C_j|_{V=0} = \frac{\epsilon_s}{W|_{V=0}} \sqrt{\frac{q\epsilon_s N_B}{2V_{bi}}} = 2.728 \times 10^{-8} \text{ F/cm}^2$$

From Eqs. 27 and 34, we obtain at a reverse bias of 4 V:

$$W|_{V=-4} \cong \sqrt{\frac{2\epsilon_s(V_{bi} - V)}{qN_D}} = \sqrt{\frac{2 \times 11.9 \times 8.85 \times 10^{-14} \times (0.906 + 4)}{1.6 \times 10^{-19} \times 8 \times 10^{15}}} = 8.99 \times 10^{-5} \text{ cm} = 0.899 \text{ }\mu\text{m}$$

$$C_j|_{V=-4} = \frac{\epsilon_s}{W|_{V=-4}} \sqrt{\frac{q\epsilon_s N_B}{2(V_{bi} - V)}} = 1.172 \times 10^{-8} \text{ F/cm}^2. \quad \blacktriangleleft$$

3.3.2 Evaluation of Impurity Distribution

The capacitance-voltage characteristics can be used to evaluate an arbitrary impurity distribution. We consider the case of p^+-n junction with a doping profile on the n -side, as shown in Fig. 12a. As before, the

incremental change in depletion layer charge per unit area dQ for an incremental change in the applied voltage dV is given by $qN(W)dW$ (i.e., the shaded area in Fig. 12b). The corresponding change in applied voltage (shaded area in Fig. 12c) is

$$dV \cong (d\mathcal{E})W \left(\frac{dQ}{\varepsilon_s} \right) W \frac{qN(W)dW^2}{2\varepsilon_s}. \quad (36)$$

By substituting W from Eq. 33, we obtain an expression for the impurity concentration at the edge of the depletion region:

$$N(W) = \frac{2}{q\varepsilon_s} \left[\frac{1}{d(1/C_j^2)/dV} \right]. \quad (37)$$

Thus, we can measure the capacitance per unit area versus reverse-bias voltage and plot $1/C_j^2$ versus V . The slope of the plot, that is, $d(1/C_j^2)/dV$, yields $N(W)$. Simultaneously, W is obtained from Eq. 33. A series of such calculations produces a complete impurity profile. This approach is referred to as the $C-V$ method for measuring impurity profiles.

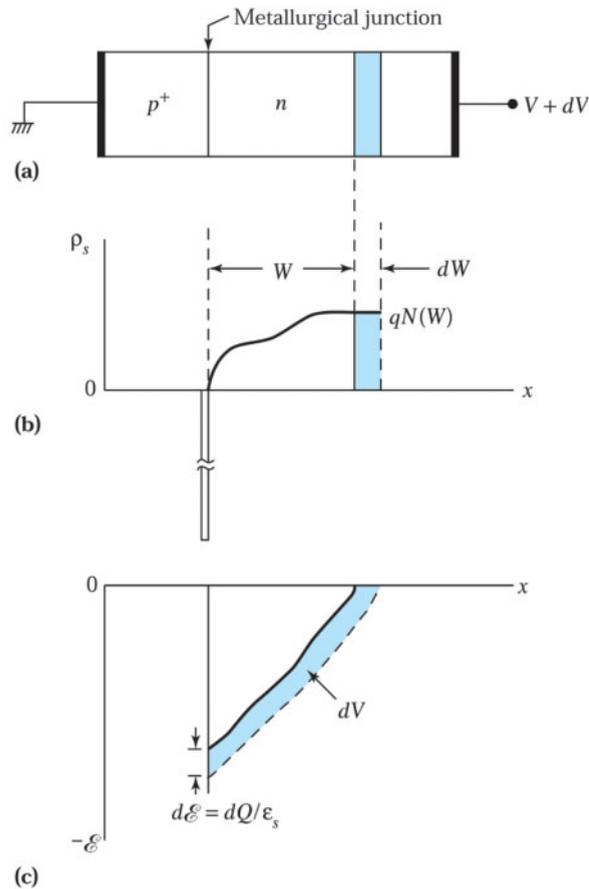


Fig. 12 (a) p^+n junction with an arbitrary impurity distribution. (b) Change in space charge distribution in the lightly doped side due to a change in applied bias. (c) Corresponding change in electric-field distribution.

For a linearly graded junction, the depletion layer capacitance is obtained from Eqs. 31 and 33:

$$C_j = \frac{\epsilon_s}{W} \left[\frac{qa\epsilon_s^2}{12(V_{bi} + V)} \right]^{1/3} \text{ F/cm}^2. \quad (38)$$

For such a junction we can plot $1/C^3$ versus V and obtain the impurity gradient and V_{bi} from the slope and the intercept, respectively.

3.3.3 Varactor

Many circuit applications employ the voltage-variable properties of reverse-biased $p-n$ junctions. A $p-n$ junction designed for such a purpose is called a *varactor*, which is a shortened form of variable reactor. As previously derived (Eq. 34 for abrupt and Eq. 38 for linearly graded junctions), the reverse-biased depletion capacitance is given by

$$C_j \propto (V_{bi} + V_R)^{-n} \quad (39)$$

or

$$C_j \propto (V_R)^{-n} \quad \text{for } V_R \gg V_{bi}, \quad (39a)$$

where $n = 1/3$ for a linearly graded junction and $n = 1/2$ for an abrupt junction. Therefore, the voltage sensitivity of C (i.e., variation of C with V_R) is greater for an abrupt junction than for a linearly graded junction. We can further increase the voltage sensitivity by using a hyperabrupt junction having an exponent n (Eq. 39) greater than $1/2$.

Figure 13 shows three p^+-n doping profiles with the donor distribution $N_D(x)$ given by $B(x/x_0)^m$, where B and x_0 are constants, $m = 1$ for a linearly graded junction, $m = 0$ for an abrupt junction, and $m = -3/2$ for a hyperabrupt junction. The hyperabrupt profile can be achieved by epitaxial growth techniques discussed in Chapter 11. To obtain the capacitance-voltage relationship, we solve the equation:

$$\frac{d^2\psi}{dx^2} = -B \left(\frac{x}{x_0} \right)^m. \quad (40)$$

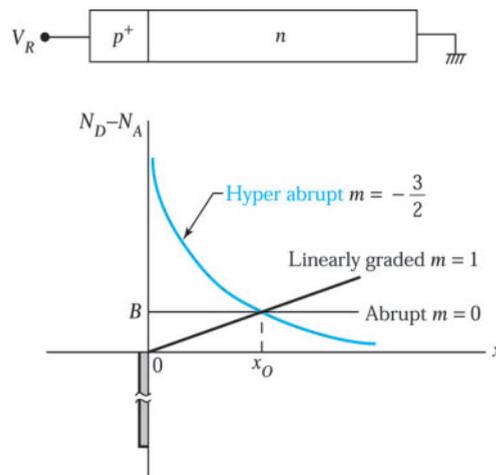


Fig. 13 Impurity profiles for hyperabrupt, one-sided abrupt, and one-sided linearly graded junctions.

Integrating Eq. 40 twice with appropriate boundary conditions gives the dependence of the depletion layer width on the reverse bias as derived for abrupt and linearly graded junctions:

$$W \propto (V_R)^{1/(m+2)}. \tag{41}$$

Therefore,

$$C_j \frac{\epsilon_s}{W} \propto (V_R)^{1/(m+2)}. \tag{42}$$

Comparing Eq. 42 with Eq. 39a yields $n = 1/(m + 2)$. For hyperabrupt junctions with $n > 1/2$, m must be a negative number.

By choosing different values for m , we can obtain a wide variety of C_j -versus- V_R dependencies for specific applications. One interesting example, shown in Fig. 13, is the case for $m = -3/2$. For this case, $n = 2$. When this varactor is connected to an inductor L in a resonant circuit, the resonant frequency varies linearly with the voltage applied to the varactor:

$$\omega_r \frac{1}{\sqrt{LC_j}} \propto \frac{1}{\sqrt{V_R^n}} V_R \quad \text{for } n = 2. \tag{43}$$

► 3.4 CURRENT-VOLTAGE CHARACTERISTICS

A voltage applied to a *p-n* junction will disturb the precise balance between the diffusion current and drift current of electrons and holes. Under forward bias, the applied voltage reduces the electrostatic potential across the depletion region, as shown in the middle of Fig. 14a. More electrons in the high-energy tail of the *n*-side conduction band shown in Fig. 22d in Chapter 1 have enough energy to surmount the smaller barrier and diffuse from the *n*-side to *p*-side. Similarly, holes in the *p*-side valence band diffuse to the *n*-side over the smaller barrier. Therefore, minority carrier injections occur, that is, electrons are injected into the *p*-side, whereas holes are injected into the *n*-side. Under reverse bias, the applied voltage increases the electrostatic potential across the depletion region, as shown in the middle of Fig. 14b. This greatly reduces the diffusion currents. For the drift current, it is almost the same despite the barrier change. Because a low concentration of minority electrons or holes in the *p* or *n* side that wander into the transition region will drift into the *n* or *p* side, the drift current depends mainly on the number of minority carriers, which travel at almost their saturation velocity. The drift current and the diffusion current coexist in the depletion region and make it more difficult to derive the current equations. Therefore, we derive the current equations only by the diffusion equations outside the depletion region. In this section, we first consider the ideal current-voltage characteristics. We then discuss departures from these ideal characteristics due to generation, recombination, and other effects.

3.4.1 Ideal Characteristics

We now derive the ideal current-voltage characteristics based on the following assumptions: (a) the depletion region has abrupt boundaries and, outside the boundaries, the semiconductor is assumed to be neutral; (b) the carrier densities at the boundaries are related by the electrostatic potential difference across the junction; (c) the low-injection condition, that is, the injected minority carrier densities are small compared with the majority carrier densities (in other words, the majority carrier densities are changed negligibly at the boundaries of neutral regions by the applied bias); (d) neither generation nor recombination current exists in the depletion region and the electron and hole currents are constant throughout the depletion region. Departures from these idealized assumptions are considered in the next section.

At thermal equilibrium, the majority carrier density in the neutral regions is essentially equal to the doping concentration. We use the subscripts *n* and *p* to denote the semiconductor type and the subscript *o* to specify the condition of thermal equilibrium. Hence, n_{no} and n_{po} are the equilibrium electron densities in the *n*- and *p*-sides, respectively. The expression for the built-in potential in Eq. 12 can be rewritten as

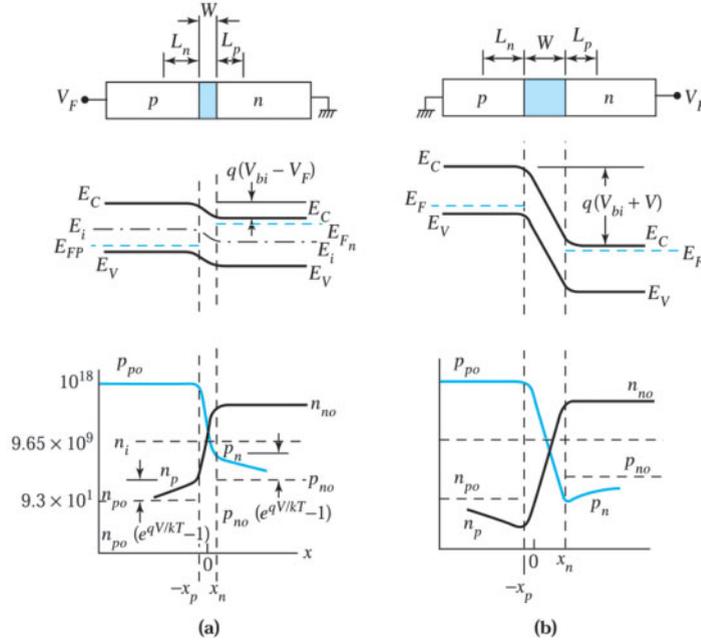


Fig. 14 Depletion region, energy band diagram and carrier distribution. (a) Forward bias. (b) Reverse bias.

$$V_{bi} = \frac{kT}{q} \ln \frac{p_{po} n_{no}}{n_i^2} = \frac{kT}{q} \ln \frac{n_{no}}{n_{po}}, \quad (44)$$

where the mass action law $p_{po} n_{po} = n_i^2$ has been used. Rearranging Eq. 44 gives

$$n_{no} = n_{po} e^{qV_{bi}/kT}. \quad (45)$$

Similarly, we have

$$p_{po} = p_{no} e^{qV_{bi}/kT}. \quad (46)$$

We note from Eqs. 45 and 46 that the electron density and the hole density at the two boundaries of the depletion region are related through the electrostatic potential difference V_{bi} at thermal equilibrium. From our second assumption we expect that the same relation holds when the electrostatic potential difference is changed by an applied voltage.

When a forward bias is applied, the electrostatic potential difference is reduced to $V_{bi} - V_F$; but when a reverse bias is applied, the electrostatic potential difference is increased to $V_{bi} + V_R$. Thus, Eq. 45 is modified to

$$n_n = n_p e^{q(V_{bi} - V)/kT}, \quad (47)$$

where n_n and n_p are the nonequilibrium electron densities at the boundaries of the depletion region in the n - and p -sides, respectively, with V positive for forward bias and negative for reverse bias. For the low-injection condition, the injected minority carrier density is much smaller than the majority carrier density; therefore, $n_n \cong n_{no}$.

Substituting this condition and Eq. 45 into Eq. 47 yields the electron density at the boundary of the depletion region on the *p*-side ($x = -x_p$):

$$n_p = n_{po} e^{qV/kT} \quad (48)$$

or

$$\frac{n_p}{n_{po}} = e^{qV/kT} \quad (48a)$$

Similarly, we have

$$p_n = p_{no} e^{qV/kT} \quad (49)$$

or

$$\frac{p_n}{p_{no}} = e^{qV/kT} \quad (49a)$$

at $x = x_n$ for the *n*-type boundary. Figure 14 shows band diagrams and carrier concentrations in a *p-n* junction under forward-bias and reverse-bias conditions. Note that the minority carrier densities at the boundaries ($-x_p$ and x_n) increase substantially above their equilibrium values under forward bias, whereas they decrease below their equilibrium values under reverse bias. Equations 48 and 49 define the minority carrier densities at the boundaries of depletion region. These equations are the most important boundary conditions for the ideal current-voltage characteristics. In the depletion region, the slopes of carrier distributions decrease with the forward bias, as shown in Fig. 14. This comes from the fast sweep of the carriers across the narrower depletion width.

Under our idealized assumptions, no current is generated within the depletion region; all currents come from the neutral regions. In the neutral *n*-region, there is no electric field, thus the steady-state continuity equation reduces to

$$\frac{d^2 p_n}{dx^2} - \frac{p_n - p_{no}}{D_p \tau_p} = 0. \quad (50)$$

The solution of Eq. 50 with the boundary conditions of Eq. 49 and $p_n(x = \infty) = p_{no}$ gives

$$p_n = p_{no} \left(e^{qV/kT} - 1 \right) e^{-(x - x_n)/L_p}, \quad (51)$$

where L_p , which is equal to $\sqrt{D_p \tau_p}$, is the diffusion length of holes (minority carriers) in the *n*-region. At $x = x_n$,

$$J_p(x_n) = qD_p \left. \frac{dp_n}{dx} \right|_{x_n} = \frac{qD_p p_{no}}{L_p} \left(e^{qV/kT} - 1 \right). \quad (52)$$

Similarly, we obtain for the neutral *p*-region

$$n_p = n_{po} \left(e^{qV/kT} - 1 \right) e^{-(x + x_p)/L_n} \quad (53)$$

and

$$J_n(x_p) = qD_n \left. \frac{dn_p}{dx} \right|_{x_p} = \frac{qD_n n_{po}}{L_n} \left(e^{qV/kT} - 1 \right), \quad (54)$$

where L_n , which is equal to $\sqrt{D_n \tau_n}$, is the diffusion length of electrons. The minority carrier densities (Eqs. 51 and 53) are shown in the middle of Fig. 15.

The graphs illustrate that the injected minority carriers recombine with the majority carriers as the minority carriers move away from the boundaries. The electron and hole currents are shown at the bottom of Fig. 15.

The hole and electron currents at the boundaries are given by Eqs. 52 and 54, respectively. The hole diffusion current will decay exponentially in the n -region with diffusion length L_p , and the electron diffusion current will decay exponentially in the p -region with diffusion length L_n .

The total current is constant throughout the device and is the sum of Eqs. 52 and 54:

$$J = J_p(x_n) + J_n(-x_p) = J_s \left(e^{qV/kT} - 1 \right), \tag{55}$$

$$J_s \equiv \frac{qD_p p_{no}}{L_p} + \frac{qD_n n_{po}}{L_n}, \tag{55a}$$

where J_s is the saturation current density. Equation 55 is the *ideal diode equation*.¹ The ideal current-voltage characteristic is shown in Figs. 16a and 16b in Cartesian and semilog plots, respectively. In the forward direction with positive bias on the p -side, for $V \geq 3kT/q$, the rate of current increase is constant, as shown in Fig. 16b. At 300 K for every decade change of current, the voltage change for an ideal diode is 60 mV ($= 2.3 kT/q$). In the reverse direction, the current density saturates at $-J_s$.

The total current for p^+n junction is

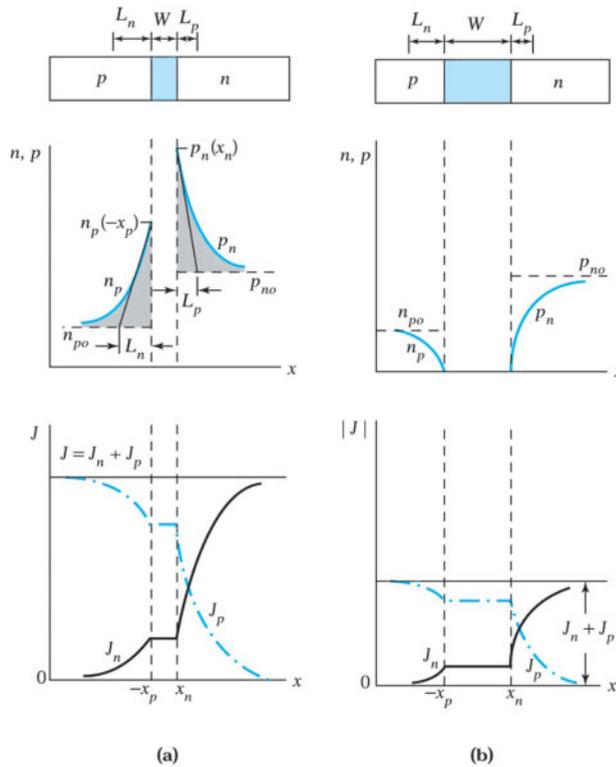


Fig. 15 Injected minority carrier distribution and electron and hole currents. (a) Forward bias. (b) Reverse bias. The figure illustrates idealized currents. In practical devices, the currents are not constant across the space charge layer.

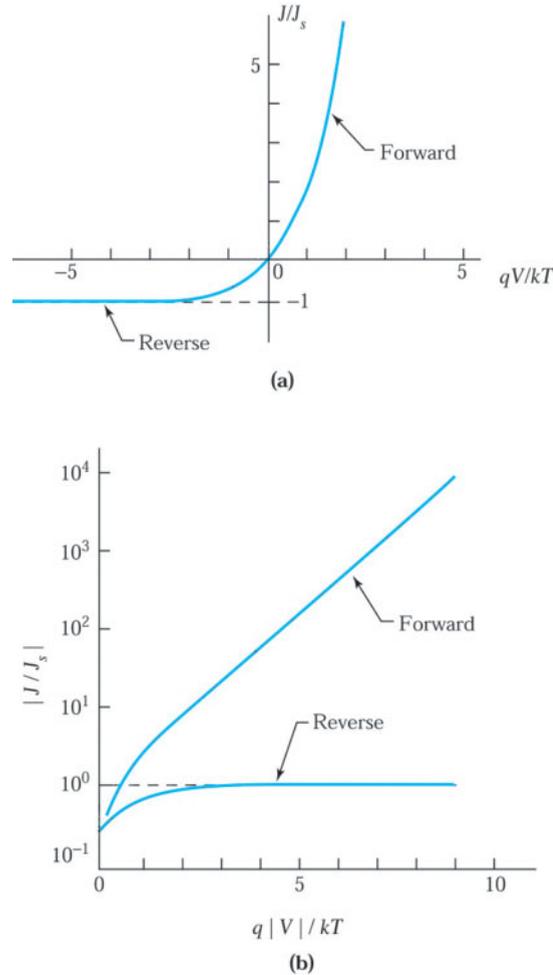


Fig. 16 Ideal current-voltage characteristics. (a) Cartesian plot. (b) Semilog plot.

$$J = \frac{qD_p p_{no}}{L_p} \left(e^{qV/kT} - 1 \right) = \frac{qD_p}{L_p} N_V \left(e^{[qV - (E_F - E_V)]/kT} - 1 \right) \quad (55b)$$

The current is small if the forward bias is less than $(E_F - E_V)/q$. The current increases rapidly if the forward bias is slightly higher than $(E_F - E_V)/q$. This is the cut-in voltage, which is slightly less than the bandgap value in electron volts, as shown in Fig. 1. Basically, the cut-in voltage increases with the bandgap.

► **EXAMPLE 5**

Calculate the ideal reverse saturation current in a Si *p-n* junction diode with a cross-sectional area of 2×10^{-4} cm². The parameters of the diode are

$$\begin{aligned} N_A &= 5 \times 10^{16} \text{ cm}^{-3}, & N_D &= 10^{16} \text{ cm}^{-3}, & n_i &= 9.65 \times 10^9 \text{ cm}^{-3} \\ D_n &= 21 \text{ cm}^2/\text{s}, & D_p &= 10 \text{ cm}^2/\text{s}, & \tau_p &= \tau_n = 5 \times 10^{-7} \text{ s}. \end{aligned}$$

SOLUTION From Eq. 55a and $L_p = \sqrt{D_p \tau_p}$, we can obtain

$$J_s = \frac{qD_p p_{n0}}{L_p} + \frac{qD_n n_{p0}}{L_n} = qn_i^2 \left(\frac{1}{N_D} \sqrt{\frac{D_p}{\tau_p}} + \frac{1}{N_A} \sqrt{\frac{D_n}{\tau_n}} \right)$$

$$1.6 \times 10^{19} \times (9.65 \times 10^9)^2 \left(\frac{1}{10^{16}} \sqrt{\frac{10}{5 \times 10^7}} + \frac{1}{5 \times 10^{16}} \sqrt{\frac{21}{5 \times 10^7}} \right)$$

$$8.58 \times 10^{12} \text{ A/cm}$$

From the cross-sectional area $A = 2 \times 10^{-4} \text{ cm}^2$, we obtain

$$I_s = A \times J_s = 2 \times 10^{-4} \times 8.58 \times 10^{12} = 1.72 \times 10^{-15} \text{ A}$$

3.4.2 Generation-Recombination and High-Injection Effects

The ideal diode equation, Eq. 55, adequately describes the current-voltage characteristics of germanium p - n junctions at low current densities. For silicon and gallium arsenide p - n junctions, however, the ideal equation can give only qualitative agreement because of the generation or recombination of carriers in the depletion region.

Consider the reverse-bias condition first. Under reverse bias, carrier concentrations in the depletion region fall far below their equilibrium concentrations. The dominant generation-recombination processes discussed in Chapter 2 are those of electron and hole emissions through bandgap generation-recombination centers. The capture processes are not important because their rates are proportional to the concentration of free carriers, which is very small in the reverse-biased depletion region.

The two emission processes operate in the steady state by alternately emitting electrons and holes. The rate of electron-hole pair generation can be obtained from Eq. 50 of Chapter 2 with the conditions $p_n < n_i$ and $n_n < n_i$:

$$G = U \left[\frac{\sigma_p \sigma_n \nu_{th} N_t}{\sigma_n \exp\left(\frac{E_t - E_i}{kT}\right) + \sigma_p \exp\left(\frac{E_i - E_t}{kT}\right)} \right] n_i \equiv \frac{n_i}{\tau_g}, \quad (56)$$

where τ_g , the generation lifetime, is the reciprocal of the expression in the square brackets. We can arrive at an important conclusion about electron-hole generation from this expression. Let us consider a simple case where $\sigma_n = \sigma_p = \sigma_o$. For this case, Eq. 56 reduces to

$$G = \frac{\sigma_o \nu_{th} N_t n_i}{2 \cosh\left(\frac{E_t - E_i}{kT}\right)}. \quad (57)$$

The generation rate reaches a maximum value at $E_t = E_i$ and falls off exponentially as E_t moves in either direction away from the middle of the bandgap. Thus, only those centers with an energy level of E_t near the intrinsic Fermi level can contribute significantly to the generation rate.

The current due to generation in the depletion region is

$$J_{gen} = \int_0^W qG dx \cong qGW = \frac{qn_i W}{\tau_g}, \quad (58)$$

where W is the depletion layer width. The total reverse current for a p^+ - n junction, that is, for $N_A \gg N_D$ and for $V_R > 3kT/q$, can be approximated by the sum of the diffusion current in the neutral regions and the generation current in the depletion region:

$$J_R \cong q \sqrt{\frac{D_p}{\tau_p} \frac{n_i^2}{N_D} + \frac{qn_i W}{\tau_g}}. \quad (59)$$

For semiconductors with large values of n_i , such as germanium, the diffusion current dominates at room temperature, and the reverse current follows the ideal diode equation. But if n_i is small, as for silicon and gallium arsenide, the generation current in the depletion region may dominate.

► **EXAMPLE 6**

Consider the Si p–n junction diode in Example 5 and assume $\tau_g = \tau_p = \tau_n$, calculate the generation current density for a reverse bias of 4 V.

SOLUTION From Eq. 20, we obtain

$$\begin{aligned} W &= \sqrt{\frac{2\epsilon_s}{q} \left(\frac{N_A + N_D}{N_A N_D} \right) (V_{bi} + V)} = \sqrt{\frac{2\epsilon_s}{q} \left(\frac{N_A + N_D}{N_A N_D} \right) \left(\frac{kT}{q} \ln \frac{N_A N_D}{n_i^2} + V \right)} \\ &= \sqrt{\frac{2 \times 11.9 \times 8.85 \times 10^{-14}}{1.6 \times 10^{-19}} \left(\frac{5 \times 10^{16} + 10^{16}}{5 \times 10^{16} \times 10^{16}} \right) \left(0.0259 \ln \frac{5 \times 10^{16} \times 10^{16}}{(9.65 \times 10^9)^2} + V \right)} \\ &= 3.97 \times \sqrt{0.758 + V} \times 10^{-5} \text{ cm.} \end{aligned}$$

Hence the generation current density is

$$\begin{aligned} J_{gen} &= \frac{qn_i W}{\tau_g} = \frac{1.6 \times 10^{-19} \times 9.65 \times 10^9}{5 \times 10^{-7}} \times 3.97 \times \sqrt{0.758 + V} \times 10^{-5} \text{ A/cm}^2 \\ &= 1.22 \times \sqrt{0.758 + V} \times 10^{-7} \text{ A/cm}^2. \end{aligned}$$

If we apply a reversed bias of 4 V, the generation current density is 2.66×10^{-7} A/cm². ◀

Under forward bias, the concentrations of both electrons and holes exceed their equilibrium values. The carriers will attempt to return to their equilibrium values by recombination. Therefore, the dominant generation-recombination processes in the depletion region are the capture processes. From Eq. 49 we obtain

$$p_n n_n \cong p_{no} n_{no} e^{qV/kT} \quad n_i^2 e^{qV/kT}. \quad (60)$$

Substituting Eq. 60 in Eq. 50 of Chapter 2 and assuming $\sigma_n = \sigma_p = \sigma_o$ yields

$$U = \frac{\sigma_o v_{th} N_t n_i^2 \left(e^{qV/kT} - 1 \right)}{n_n + p_n + 2n_i \cosh \frac{E_i - E_t}{kT}}. \quad (61)$$

In either recombination or generation, the most effective centers are those located near E_i . As practical examples, gold and copper yield effective generation-recombination centers in silicon where the values of $E_i - E_t$ are 0.02 V for gold and –0.02 eV for copper. In gallium arsenide, chromium gives an effective center with an $E_i - E_t$ value of 0.08 eV.

Equation 61 can be simplified for the case $E_i = E_t$:

$$U = \sigma_o v_{th} N_t \frac{n_i^2 \left(e^{qV/kT} - 1 \right)}{n_n + p_n + 2n_i} \quad (62)$$

For a given forward bias, U reaches its maximum value at a location in the depletion region either where the denominator $n_n + p_n + 2n_i$ is a minimum or where the sum of the electron and hole concentrations, $n_n + p_n$, is at its minimum value. Since the product of these concentrations is a constant given by Eq. 60, the condition $d(p_n + n_n) = 0$ leads to

$$dp_n = -dn_n \frac{p_n n_n}{p_n^2} dp_n \quad (63)$$

or

$$p_n = n_n \quad (64)$$

as the condition for the minimum. This condition exists at the location in the depletion region, where E_i is halfway between E_{Fp} and E_{Fn} , as illustrated in the middle of Fig. 14a. Here, the carrier concentrations are

$$p_n = n_n = n_i e^{qV/2kT} \quad (65)$$

and therefore

$$U_{\max} = \sigma_o \nu_{th} N_i \frac{n_i^2 (e^{qV/kT} - 1)}{2n_i (e^{qV/2kT} + 1)}. \quad (66)$$

For $V > 3kT/q$,

$$U_{\max} \cong \frac{1}{2} \sigma_o \nu_{th} N_i n_i e^{qV/2kT}. \quad (67)$$

The recombination current is then

$$J_{rec} = \int_0^W qU dx \cong \frac{qW}{2} \sigma_o \nu_{th} N_i n_i e^{qV/2kT} \frac{qW n_i}{2\tau_r} e^{qV/2kT}, \quad (68)$$

where τ_r is the effective recombination lifetime given by $1/(\sigma_o \nu_{th} N_i)$. The total forward current can be approximated by the sum of Eqs. 55 and 68. For $p_{no} \gg n_{po}$ and $V > 3kT/q$ we have

$$J_F = q \sqrt{\frac{D_p}{\tau_p} \frac{n_i^2}{N_D} e^{qV/kT} + \frac{qW n_i}{2\tau_r} e^{qV/2kT}}. \quad (69)$$

In general, the experimental results can be represented empirically by

$$J_F \approx \exp\left(\frac{qV}{\eta kT}\right), \quad (70)$$

where the factor η is called the *ideality factor*. When the ideal diffusion current dominates, $\eta = 1$; whereas when the recombination current dominates, $\eta = 2$. When the two currents are comparable, η has a value between 1 and 2.

Figure 17 shows the measured forward characteristics of a silicon and gallium arsenide p - n junction at room temperature.² At low current levels, recombination current dominates and $\eta = 2$. At higher current levels, diffusion current dominates and η approaches 1.

At even higher current levels, we notice that the current departs from the ideal $\eta = 1$ situation and increases more gradually with forward voltage. This phenomenon is associated with two effects: series resistance and high injection. We first consider the series resistance effect. At both low- and medium-current levels, the IR drop across the neutral regions is usually small compared with kT/q (26 mV at 300 K), where I is the forward current and R is the series resistance. For example, for a silicon diode with $R = 1.5$ ohms, the IR drop at 1 mA is only 1.5 mV. However, at 100 mA, the IR drop becomes 0.15 V, which is six times larger than kT/q . This IR drop reduces the bias across the depletion region; therefore, the current becomes

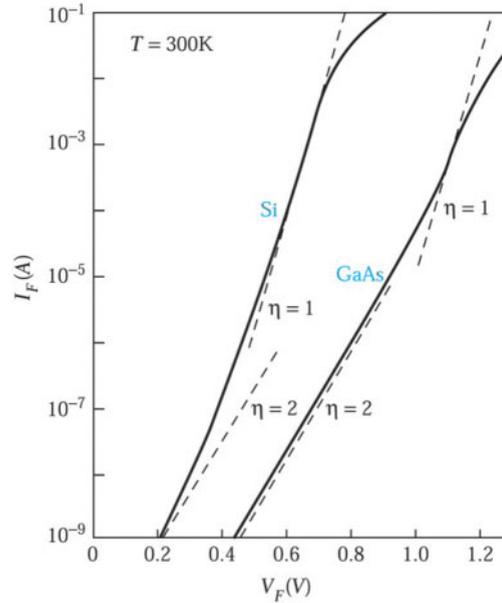


Fig. 17 Comparison of the forward current-voltage characteristics of Si and GaAs diodes² at 300 K. Dashed lines indicate slopes of different ideality factors η .

$$I \cong I_S \exp\left[\frac{q(V - IR)}{kT}\right] = \frac{I_S \exp(qV / kT)}{\exp\left[\frac{q(IR)}{kT}\right]} \quad (71)$$

and the ideal diffusion current is reduced by the factor $\exp[q(IR)/kT]$.

At high-current densities, the injected minority carrier density is comparable to the majority concentration; that is, at the n -side of the junction, $p_n(x = x_n) \cong n_n$. This is the high-injection condition. By substituting the high-injection condition in Eq. 60, we obtain $p_n(x = x_n) \cong n_i \exp(qV/2kT)$. Using this as a boundary condition, the current becomes roughly proportional to $\exp(qV/2kT)$. Thus, the current increases at a slower rate under the high-injection condition.

3.4.3 Temperature Effect

Operating temperature has a profound effect on device performance. In both the forward-bias and reverse-bias conditions, the magnitudes of the diffusion and the recombination-generation currents depend strongly on temperature. We consider the forward-bias case first. The ratio of hole diffusion current to the recombination is given by

$$\frac{I_{diffusion}}{I_{recombination}} \cong 2 \frac{n_i}{N_D} \frac{L_p}{W} \frac{\tau_r}{\tau_p} e^{qV/2kT} \approx \exp\left(\frac{E_g - qV}{2kT}\right). \quad (72)$$

This ratio depends on both the temperature and the semiconductor bandgap. Figure 18a shows the temperature dependence of the forward characteristics of a silicon diode. At room temperature for small forward voltages, the recombination current generally dominates, whereas at higher forward voltages the diffusion current usually dominates. At a given forward bias, as the temperature increases, the diffusion current will increase more rapidly than the recombination current. Therefore, the ideal diode equation will be followed over a wide range of forward biases as the temperature increases.

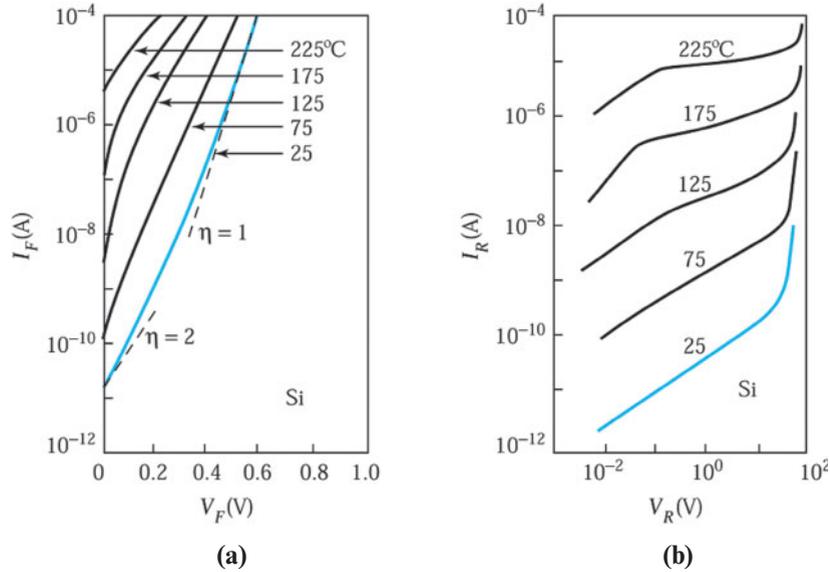


Fig. 18 Temperature dependence of the current-voltage characteristics of a Si diode². (a) Forward bias. (b) Reverse bias.

The temperature dependence of the saturation current density J_s (Eq. 55a) for a one-sided p^+n junction in which diffusion current dominates is given by

$$J_s \cong \frac{qD_p P_{no}}{L_p} \approx n_i^2 \approx \exp\left(-\frac{E_g}{kT}\right). \quad (73)$$

Thus, the activation energy obtained from the slope of a plot of J_s versus $1/T$ corresponds to the energy bandgap E_g .

In the reverse-bias condition for a p^+n junction, the ratio of the diffusion current to the generation current is

$$\frac{I_{diffusion}}{I_{generation}} = \frac{n_i}{N_D} \frac{L_p}{W} \frac{\tau_g}{\tau_p}. \quad (74)$$

This ratio is proportional to the intrinsic carrier density n_i . As the temperature increases, the diffusion current eventually dominates. Figure 18b shows the effects of temperature on the reverse characteristics of a silicon diode. At low temperatures, the generation current dominates and the reverse current varies as $\sqrt{V_R}$ in accordance with Eq. 58 for an abrupt junction (i.e., $W \sim \sqrt{V_R}$). As the temperature increases beyond 175°C, the current demonstrates a saturation tendency for $V_R \geq 3kT/q$, at which point the diffusion current becomes dominant.

► 3.5 CHARGE STORAGE AND TRANSIENT BEHAVIOR

Under forward bias, electrons are injected from the n -region into the p -region and holes are injected from the p -region into the n -region. Once injected across the junction, the minority carriers recombine with the majority carriers and decay exponentially with distance, as shown in Fig. 15a. These minority-carrier distributions lead to current flow and to charge storage in the p - n junction. We consider the stored charge, its effect on junction capacitance, and the transient behavior of the p - n junction due to sudden changes of bias.

3.5.1 Minority-Carrier Storage

The charge of injected minority carriers per unit area stored in the neutral n -region can be found by integrating the excess holes in the neutral region, shown as the shaded area in the middle of Fig. 15a, using Eq. 51:

$$\begin{aligned}
 Q_p &= q \int_{x_n}^{\infty} (p_n - p_{no}) dx \\
 &= q \int_{x_n}^{\infty} p_{no} (e^{qV/kT} - 1) e^{-(x - x_n)/L_p} dx \\
 &= qL_p p_{no} (e^{qV/kT} - 1).
 \end{aligned} \tag{75}$$

L_p is the average distance of a hole diffusion before recombining. The stored charge can be regarded as the hole diffusion with an average distance of L_p away from the boundary of the depletion region. The number of stored minority carriers depends on both the diffusion length and the charge density at the boundary of the depletion region. A similar expression can be obtained for the stored electrons in the neutral p -region. We can express the stored charge in terms of the injected current. From Eqs. 52 and 75, we have

$$Q_p = \frac{L_p^2}{D_p} J_p(x_n) = \tau_p J_p(x_n). \tag{76}$$

The average lifetime of holes in n -side is τ_p . Thus, the stored charges Q_p must be replenished every τ_p seconds. Equation 76 states that the amount of stored charge depends on the current and lifetime of the minority carriers.

► **EXAMPLE 7**

For an ideal abrupt silicon p^+n junction with $N_D = 8 \times 10^{15} \text{ cm}^{-3}$, calculate the stored minority carriers per unit area in the neutral n -region when a forward bias of 1V is applied. The diffusion length of the holes is 5 μm .

SOLUTION From Eq. 75, we obtain

$$\begin{aligned}
 Q_p &= qL_p p_{no} (e^{qV/kT} - 1) = 1.6 \times 10^{-19} \times 5 \times 10^{-4} \times \frac{(9.65 \times 10^9)^2}{8 \times 10^{15}} \times (e^{\frac{1}{0.0259}} - 1) \\
 &= 4.69 \times 10^{-2} \text{ C/cm}^2.
 \end{aligned}$$

3.5.2 Diffusion Capacitance

The depletion-layer capacitance considered previously accounts for most of the junction capacitance when the junction is reverse biased. When the junction is forward biased, there is an additional significant contribution to junction capacitance from the rearrangement of the stored charges in the neutral regions. This is called the *diffusion capacitance*, denoted C_d , a term derived from the ideal-diode case in which minority carriers move across the neutral region by diffusion.

The diffusion capacitance of the stored holes in the neutral n -region is obtained by applying the definition $C_d \equiv AdQ_p/dV$ to Eq. 75:

$$C_d = \frac{Aq^2 L_p p_{no}}{kT} e^{qV/kT}, \tag{77}$$

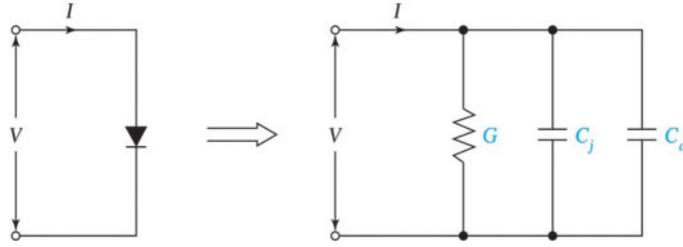


Fig. 19 Small-signal equivalent circuit of a p - n junction.

where A is the device cross-section area. We may add the contribution to C_d of the stored electrons in the neutral p -region in cases of significant storage. For a p^+ - n junction, however, $n_{po} \ll p_{no}$, and the contribution to C_d of the stored electrons becomes insignificant. Under reverse bias (i.e., V is negative), Eq. 77 shows that C_d is inconsequential because of negligible minority-carrier storage.

In many applications we prefer to represent a p - n junction by an equivalent circuit. In addition to diffusion capacitance C_d and depletion capacitance C_j , we must include conductance to account for the current through the device. In the ideal diode the conductance can be obtained from Eq. 55:

$$G = \frac{AdJ}{dV} = \frac{qA}{kT} J_s e^{qV/kT} = \frac{qA}{kT} (J + J_s) \cong \frac{qI}{kT}. \quad (78)$$

The diode equivalent circuit is shown in Fig. 19, where C_j stands for the total depletion capacitance (i.e., the result in Eq. 33 times the device area A). For low-voltage, sinusoidal excitation of a diode that is biased quiescently (i.e., at dc), the circuit shown in Fig. 19 provides adequate accuracy. Therefore, we refer to it as the diode small-signal equivalent circuit.

3.5.3 Transient Behavior

For switching applications, the forward-to-reverse-bias transition must be nearly abrupt and the transient time should be short. Figure 20a shows a simple circuit where a forward current I_F flows through a p - n junction. At time $t = 0$, switch S is suddenly thrown to the right and an initial reverse current $I_R \cong V/R$ flows. The transient time t_{off} , plotted in Fig. 20b, is the time required for the current to reach 10% of the initial reverse current I_R .

The transient time may be estimated as follows. Under the forward-bias condition, the stored minority carriers in the n -region for a p^+ - n junction is given by Eq. 76:

$$Q_p = \tau_p J_p = \tau_p \frac{I_F}{A}, \quad (79)$$

where I_F is the total forward current and A is the device area. If the average current flowing during the turn-off period is $I_{R,ave}$, the turn-off time is the length of time required to remove the total stored charge Q_p :

$$t_{off} \cong \frac{Q_p A}{I_{R,ave}} = \tau_p \left(\frac{I_F}{I_{R,ave}} \right). \quad (80)$$

Thus the turn-off time depends on both the ratio of forward to reverse currents and the lifetime of the minority carriers. The result of a more precise turn-off time calculation³ taking into account the time-dependent minority-carrier diffusion problem is shown in Fig. 21. For fast-switching devices, we must reduce the lifetime of the minority carriers. Therefore, we usually introduce recombination-generation centers that have energy levels located near mid-bandgap, such as gold in silicon.

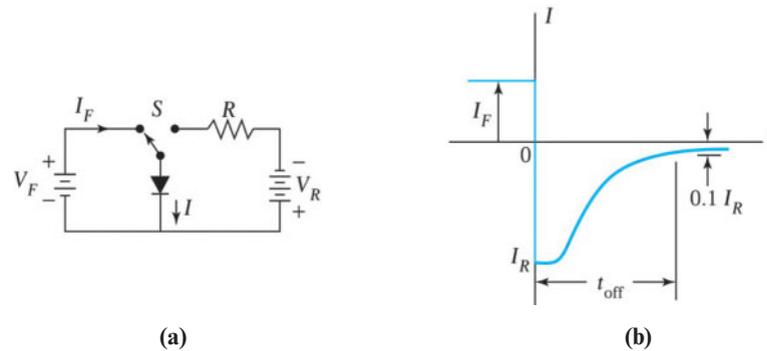


Fig. 20 Transient behavior of a *p-n* junction. (a) Basic switching circuit. (b) Transient response of the current switched from forward bias to reverse bias.

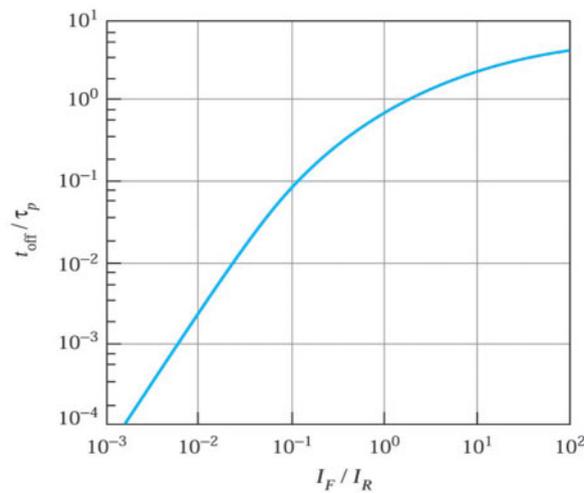


Fig. 21 Normalized transient time versus the ratio of forward current to reverse current.³

► 3.6 JUNCTION BREAKDOWN

When a sufficiently large reverse voltage is applied to a *p-n* junction, the junction breaks down and conducts a very large current. Although the breakdown process is not inherently destructive, the maximum current must be limited by an external circuit to avoid excessive junction heating. Two important breakdown mechanisms are the tunneling effect and avalanche multiplication. We consider the first mechanism briefly and then discuss avalanche multiplication in detail, because avalanche breakdown imposes an upper limit on the reverse bias for most diodes. Avalanche breakdown also limits the collector voltage of a bipolar transistor (Chapter 4) and the drain voltage of a MOSFET (Chapters 5 and 6). In addition, the avalanche multiplication mechanisms can generate microwave power, as in an IMPATT diode (Chapter 8), and detect optical signals, as in an avalanche photodetector (Chapter 10).

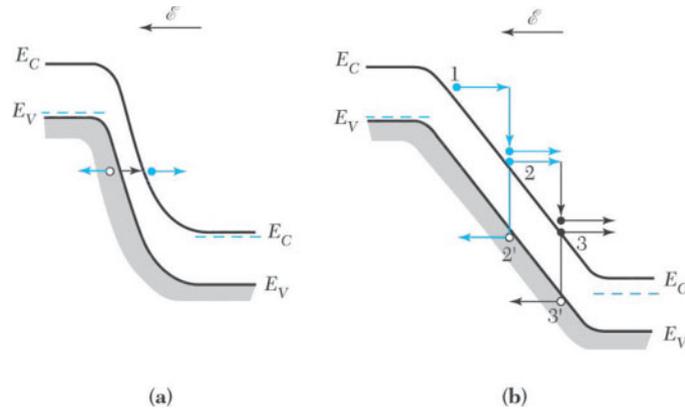


Fig. 22 Energy band diagrams under junction-breakdown conditions. (a) Tunneling effect. (b) Avalanche multiplication.

3.6.1 Tunneling Effect

When a high electric field is applied to a p - n junction in the reverse direction, a valence electron can make a transition from the valence band to the conduction band, as shown in Fig. 22a. This process, in which an electron penetrates through the energy bandgap, is called tunneling.

The tunneling process is discussed in Chapter 2. Tunneling occurs only if the electric field is very high. The typical field for silicon and gallium arsenide is about 10^6 V/cm or higher. To achieve such a high field, the doping concentrations for both p - and n -regions must be quite high ($> 5 \times 10^{17}$ cm $^{-3}$). The breakdown mechanisms for silicon and gallium arsenide junctions with breakdown voltages of less than about $4E_g/q$, where E_g is the bandgap, are the result of the tunneling effect. For junctions with breakdown voltages in excess of $6E_g/q$, the breakdown mechanism is the result of avalanche multiplication. At voltages between 4 and $6E_g/q$, the breakdown is due to a mixture of both avalanche multiplication and tunneling.⁴

3.6.2 Avalanche Multiplication

The avalanche multiplication process is illustrated in Fig. 22b. The p - n junction, such as a p^+ - n one-sided abrupt junction with a doping concentration of $N_D \cong 10^{17}$ cm $^{-3}$ or less, is under reverse bias. This figure is essentially the same as Fig. 26 in Chapter 2. A thermally generated electron in the depletion region (designated by 1) gains kinetic energy from the electric field. If the field is sufficiently high, the electron can gain enough kinetic energy that on collision with an atom, it can break the lattice bonds, creating an electron-hole pair (2 and 2'). The newly created electron and hole both acquire kinetic energy from the field and create additional electron-hole pairs (e.g., 3 and 3'). These in turn continue the process, creating other electron-hole pairs. This process is therefore called *avalanche multiplication*.

To derive the breakdown condition, we assume that a current I_{no} is incident at the left-hand side of the depletion region of width W , as shown in Fig. 23. If the electric field in the depletion region is high enough to initiate the avalanche multiplication process, the electron current I_n will increase with distance through the depletion region to reach a value $M_n I_{no}$ at W , where M_n , the multiplication factor, is defined as

$$M_n \equiv \frac{I_n(W)}{I_{no}}. \quad (81)$$

Similarly, the hole current I_p increases from $x = W$ to $x = 0$. The total current $I = (I_p + I_n)$ is constant at steady state. The incremental electron current at x equals the number of electron-hole pairs generated per second in the distance dx :

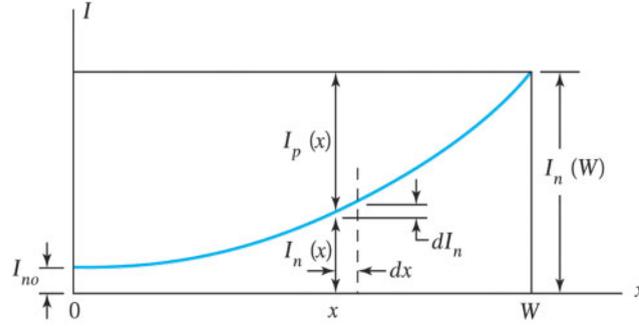


Fig. 23 Depletion region in a p-n junction with multiplication of an incident current.

$$d\left(\frac{I_n}{q}\right) = \left(\frac{I_n}{q}\right)(\alpha_n dx) + \left(\frac{I_p}{q}\right)(\alpha_p dx) \quad (82)$$

or

$$\frac{dI_n}{dx} + (\alpha_p - \alpha_n)I_n = \alpha_p I, \quad (82a)$$

where α_n and α_p are the electron and hole ionization rates, respectively. If we use the simplified assumption that $\alpha_n = \alpha_p = \alpha$, the solution of Eq. 82a is

$$\frac{I_n(W) - I_n(0)}{I} = \int_0^W \alpha dx. \quad (83)$$

From Eqs. 81 and 83, we have

$$1 - \frac{1}{M_n} = \int_0^W \alpha dx. \quad (83a)$$

The avalanche breakdown voltage is defined as the voltage at which M_n approaches infinity. Hence, the breakdown condition is given by

$$\int_0^W \alpha dx = 1. \quad (84)$$

From both the breakdown condition described above and the field dependence of the ionization rates, we may calculate the critical field (i.e., the maximum electric field at breakdown) at which the avalanche process takes place. Using measured α_n and α_p (Fig. 27 in Chapter 2), the critical field \mathcal{E}_c is calculated for silicon and gallium arsenide one-sided abrupt junctions and shown in Fig. 24 as functions of the impurity concentration of the substrate. Also indicated is the critical field for the tunneling effect. It is evident that tunneling occurs only in semiconductors having high doping concentrations.

With the critical field determined, we may calculate the breakdown voltages. As discussed previously, voltages in the depletion region are determined from the solution of Poisson's equation:

$$V_B(\text{breakdown voltage}) = \frac{\mathcal{E}_c W}{2} = \frac{\epsilon_s \mathcal{E}_c^2}{2q} (N_B)^{-1} \quad (85)$$

for one-sided abrupt junctions and

$$V_B = \frac{2\epsilon_c W}{3} = \frac{4\epsilon_c^{3/2}}{3} \left(\frac{2\epsilon_s}{q} \right)^{1/2} (a)^{-1/2} \tag{86}$$

for linearly graded junctions, where N_B is the background doping of the lightly doped side, ϵ_s is the semiconductor permittivity, and a is the impurity gradient. Since the critical field is a slowly varying function of either N_B or a , the breakdown voltage, as a first-order approximation, varies as N_B^{-1} for abrupt junctions and as $a^{-1/2}$ for linearly graded junctions.

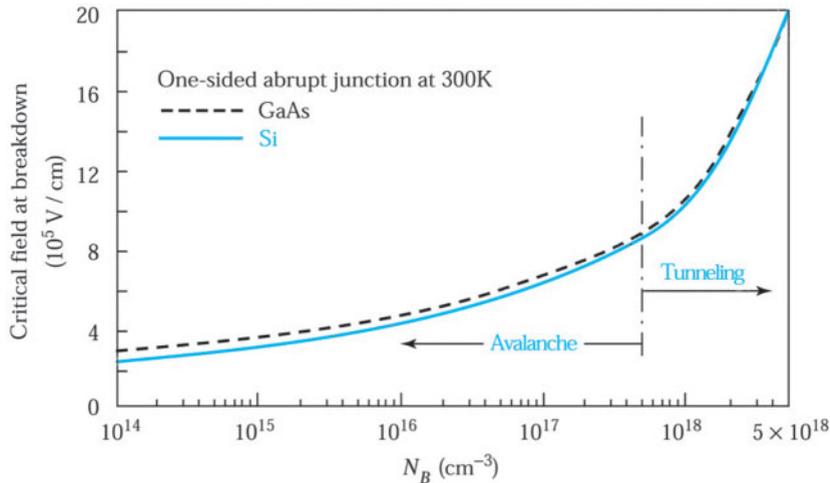


Fig. 24 Critical field at breakdown versus background doping for Si and GaAs one-sided abrupt junctions.⁵

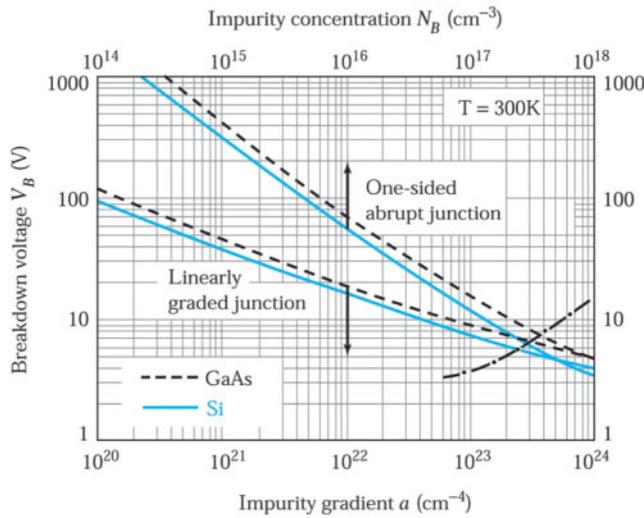


Fig. 25 Avalanche breakdown voltage versus impurity concentration for a one-sided abrupt junction and avalanche breakdown voltage versus impurity gradient for a linearly graded junction in Si and GaAs. Dash-dot line indicates the onset of the tunneling mechanism.⁵

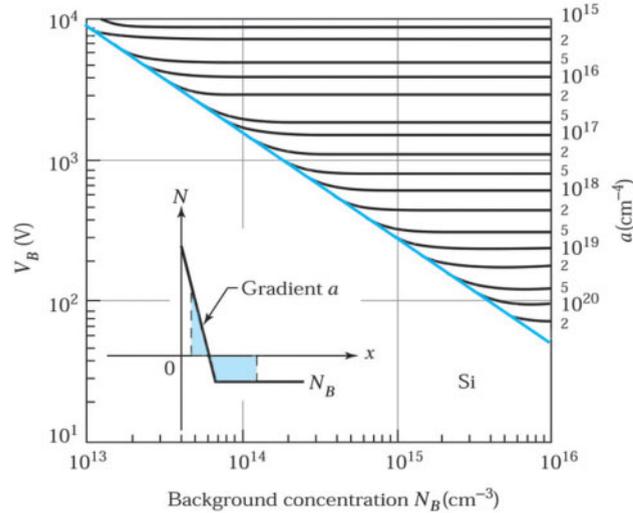


Fig. 26 Breakdown voltage for diffused junctions. Inset shows the space charge distribution.⁶

Figure 25 shows the calculated avalanche breakdown voltages for silicon and gallium arsenide junctions.⁵ The dash-dot line (to the right) at high dopings or high-impurity gradients indicates the onset of the tunneling effect. Gallium arsenide has higher breakdown voltages than silicon for a given N_B or a , mainly because of its larger bandgap. The larger the bandgap, the larger the critical field must be for sufficient kinetic energy to be gained between collisions. As Eqs. 85 and 86 demonstrate, the larger critical field, in turn, gives rise to higher breakdown voltage.

The inset in Fig. 26 shows the space-charge distribution of a diffused junction with a linear gradient near the surface and a constant doping inside the semiconductor. The breakdown voltage lies between the two limiting cases of abrupt junction and linearly graded junction considered previously.⁶ For large a and low N_B , the breakdown voltage of the diffused junctions is given by the abrupt junction results shown by the bottom line in Fig. 26, whereas for small a and high N_B , V_B is given by the linearly graded junction results indicated by the parallel lines in Fig. 26.

▶ EXAMPLE 8

Calculate the breakdown voltage for a Si one-sided $p^+ - n$ abrupt junction with $N_D = 5 \times 10^{16} \text{ cm}^{-3}$.

SOLUTION From Fig. 24, we see that the critical field at breakdown for a Si one-sided abrupt junction is about $5.7 \times 10^5 \text{ V/cm}$. Then from Eq. 85, we obtain

$$\begin{aligned} V_B(\text{breakdown voltage}) &= \frac{\epsilon_c W}{2} = \frac{\epsilon_s \epsilon_c^2}{2q} (N_B)^{-1} \\ &= \frac{11.9 \times 8.85 \times 10^{-14} \times (5.7 \times 10^5)^2}{2 \times 1.6 \times 10^{-19}} (5 \times 10^{16})^{-1} \\ &= 21.4 \text{ V} \end{aligned}$$

In Figs. 25 and 26 we assume that the semiconductor layer is thick enough to support the reverse-biased depletion layer width W_m at breakdown. If the semiconductor layer W is smaller than W_m , as shown in the inset of Fig. 27, the device will be punched through; that is, the depletion layer will reach the $n - n^+$ interface prior to breakdown. Increase the reverse bias further and the device will break down. The critical field ϵ_c is essentially the same as that shown in Fig. 24. Therefore, the breakdown voltage V'_B for the punch-through diode is

$$\frac{V'_B}{V_B} = \frac{\text{shaded area in Fig. 27 inset}}{(\epsilon_s W_m)/2} = \left(\frac{W}{W_m}\right)\left(2 - \frac{W}{W_m}\right), \quad (87)$$

Punch-through occurs when the doping concentration N_B becomes sufficiently low, as in a $p^+-\pi-n^+$ or p^+-v-n^+ diode, where π stands for a lightly doped p -type and v stands for a lightly doped n -type semiconductor. The breakdown voltages for such diodes calculated from Eqs. 85 and 87 are shown in Fig. 27. For a given thickness, the breakdown voltage approaches a constant value as the doping decreases.

► **EXAMPLE 9**

For a GaAs p^+-n one-sided abrupt junction with $N_D = 8 \times 10^{14} \text{ cm}^{-3}$, calculate the depletion width at breakdown. If the n -type region of this structure is reduced to $20 \mu\text{m}$, calculate the breakdown voltage.

SOLUTION From Fig. 25, we can find that the breakdown voltage (V_B) is about 500 V, which is much larger than the built-in voltage (V_{bi}). And from Eq. 27, we obtain

$$W \sqrt{\frac{2\epsilon_s(V_{bi} + V)}{qN_B}} \cong \sqrt{\frac{2 \times 12.4 \times 8.85 \times 10^{-14} \times 500}{1.6 \times 10^{-19} \times 8 \times 10^{14}}} = 2.93 \times 10^3 = 29.3 \mu\text{m}.$$

When the n -type region reduces to $20 \mu\text{m}$, punch-through will occur first. From Eq. 87, we can obtain

$$\frac{V'_B}{V_B} = \frac{\text{shaded area in Fig. 27 inset}}{(\epsilon_s W_m)/2} = \left(\frac{W}{W_m}\right)\left(2 - \frac{W}{W_m}\right),$$

$$V'_B = V_B \left(\frac{W}{W_m}\right)\left(2 - \frac{W}{W_m}\right) = 500 \times \left(\frac{20}{29.3}\right)\left(2 - \frac{20}{29.3}\right) = 449 \text{ V}.$$

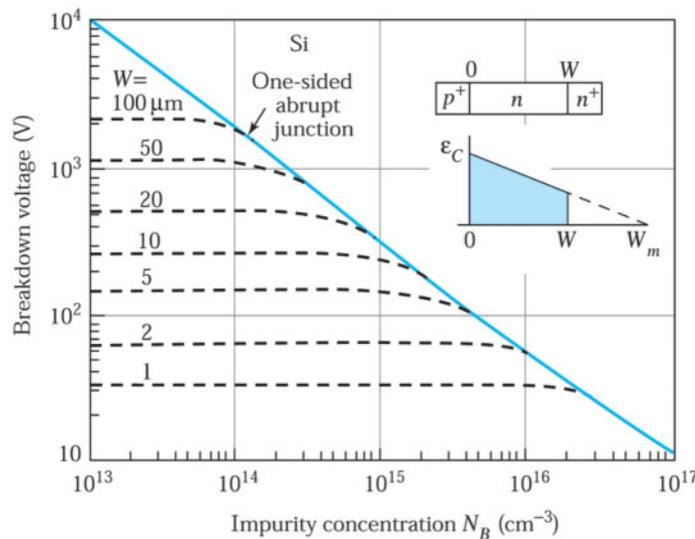


Fig. 27 Breakdown voltage for $p^+-\pi-n^+$ and p^+-v-n^+ junctions. W is the thickness of the lightly doped p -type (π) or the lightly doped n -type (v) region.

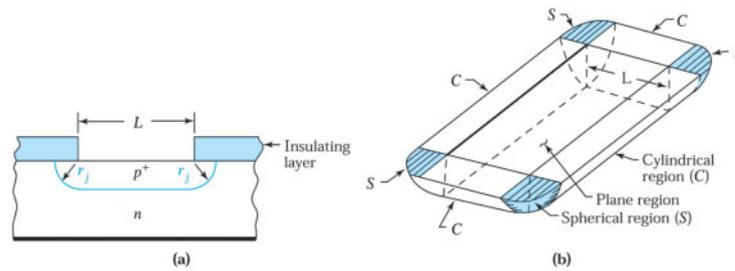


Fig. 28 (a) Planar diffusion process that forms junction curvature near the edge of the diffusion mask, where r_j is the radius of curvature. (b) Cylindrical and spherical regions formed by diffusion through a rectangular mask.

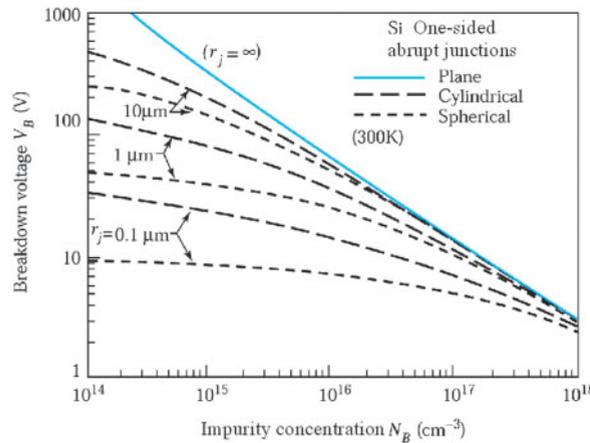


Fig. 29 Breakdown voltage versus impurity concentration for a one-sided abrupt doping profile with cylindrical and spherical junction geometries,⁷ where r_j is the radius of curvature indicated in Fig. 28.

Another important consideration of breakdown voltage is the junction curvature effect.⁷ When a *p-n* junction is formed by diffusion through a window in the insulating layer on a semiconductor, the impurities diffuse downward and sideways (see Chapter 14). Hence, the junction has a plane (or flat) region with nearly cylindrical edges, as shown in Fig. 28a. If the diffusion mask contains sharp corners, the corner of the junction will acquire the roughly spherical shape shown in Fig. 28b. Because the spherical or cylindrical regions of the junction have higher field intensity, they determine the avalanche breakdown voltage. The calculated results for silicon one-sided abrupt junctions are shown in Fig. 29. The solid line represents the plane junctions considered previously. Note that as the junction radius r_j becomes smaller, the breakdown voltage decreases dramatically, especially for spherical junctions at low impurity concentrations.

► 3.7 HETEROJUNCTION

A heterojunction is defined as a junction formed between two dissimilar semiconductors. Figure 30a shows the energy band diagram of two isolated pieces of semiconductors prior to the formation of a heterojunction. The two semiconductors are assumed to have different energy bandgaps E_g , different dielectric permittivities ϵ_s , different work functions $q\phi_s$, and different electron affinities $q\chi$. The work function is defined as the energy required to remove an electron from Fermi level E_F to a position just outside the material (the vacuum level). The electron affinity is the energy required to remove an electron from the bottom of the conduction band E_C to the vacuum level. The difference in energy of the conduction band edges in the two semiconductors is represented by ΔE_C , and the difference in energy of the valence band edges is represented by ΔE_V . From Fig. 30a, ΔE_C and ΔE_V can be expressed by

$$\Delta E_C = q(\chi_2 - \chi_1) \tag{88a}$$

and

$$\Delta E_V = E_{g1} + qx_1 - (E_{g2} + qx_2) = \Delta E_g - \Delta E_C \tag{88b}$$

where ΔE_g is the energy band difference and $\Delta E_g = E_{g1} - E_{g2}$.

Figure 30b shows the equilibrium band diagram of an ideal abrupt heterojunction formed between these semiconductors.⁸ In this diagram it is assumed that there is a negligible number of traps or generation-recombination centers at the interface of the two dissimilar semiconductors. Note that this assumption is valid only when heterojunctions are formed between semiconductors with closely matched lattice constants. Therefore, we must choose lattice-matched materials to satisfy the assumption.⁸ For example, the $\text{Al}_x\text{Ga}_{1-x}\text{As}$ materials, with x from 0 to 1, are among the most important materials for heterojunctions. When $x = 0$, we have GaAs, with a

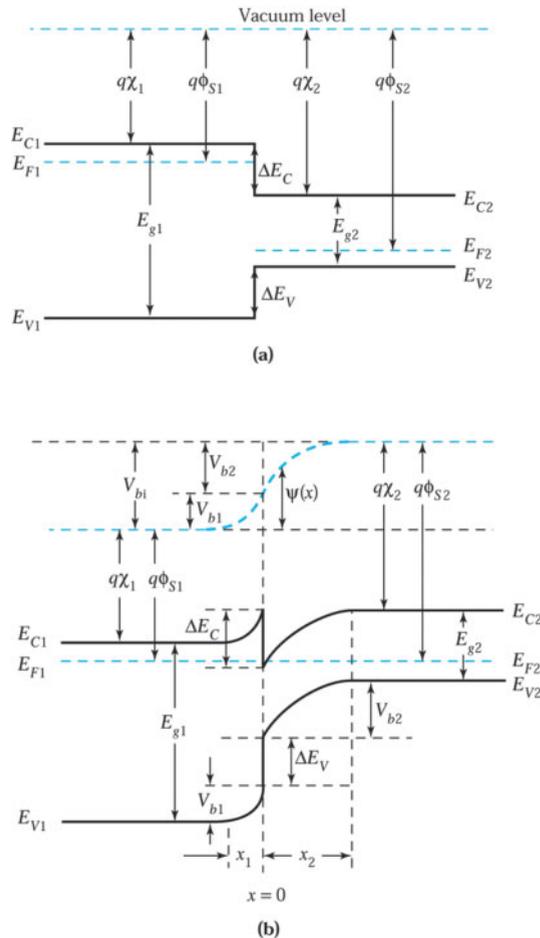


Fig. 30 (a) Energy band diagram of two isolated semiconductors. (b) Energy band diagram of an ideal $n-p$ heterojunction at thermal equilibrium.

⁸Lattice-mismatched epitaxy, also called strained-layer epitaxy, is considered in Section 11.6.

bandgap of 1.42 eV and a lattice constant of 5.6533 Å at 300 K. When $x = 1$, we have AlAs, with a bandgap of 2.17 eV and a lattice constant of 5.6605 Å. The bandgap for the ternary $\text{Al}_x\text{Ga}_{1-x}\text{As}$ increases with x ; however, the lattice constant remains essentially constant. Even for the extreme cases where $x = 0$ and $x = 1$, the lattice constant mismatch is only 0.1%.

There are two basic requirements in the construction of the energy band diagram: (a) the Fermi level must be the same on both sides of the interface in thermal equilibrium, and (b) the vacuum level must be continuous and parallel to the band edges. Because of these requirements, the discontinuity in conduction band edges ΔE_c and valence band edges ΔE_v will be unaffected by doping as long as the bandgap E_g and electron affinity χ are not functions of doping (i.e., as in nondegenerate semiconductors). The total built-in potential V_{bi} can be expressed by

$$V_{bi} = V_{b1} + V_{b2}, \quad (89)$$

where V_{b1} and V_{b2} are the electrostatic potentials at equilibrium in semiconductors 1 and 2, respectively.

Under the conditions that the potential and the *free-carrier flux density* (defined as the rate of free-carrier flow through a unit area) are continuous at the heterointerface, we can derive the depletion widths and capacitance from the Poisson equation using the conventional depletion approximation. One boundary condition is the continuity of electric displacement, that is, $\epsilon_1 \mathcal{E}_1 = \epsilon_2 \mathcal{E}_2$, where \mathcal{E}_1 and \mathcal{E}_2 are the electric fields at the interface ($x = 0$) in semiconductors 1 and 2, respectively. V_{b1} and V_{b2} are given by

$$V_{b1} = \frac{\epsilon_2 N_2 (V_{bi} - V)}{\epsilon_1 N_1 + \epsilon_2 N_2}, \quad (90a)$$

$$V_{b2} = \frac{\epsilon_1 N_1 (V_{bi} - V)}{\epsilon_1 N_1 + \epsilon_2 N_2}, \quad (90b)$$

where N_1 and N_2 are the doping concentrations in semiconductors 1 and 2, respectively. The depletion widths x_1 and x_2 can be obtained by

$$x_1 = \sqrt{\frac{2\epsilon_1 \epsilon_2 N_2 (V_{bi} - V)}{q N_1 (\epsilon_1 N_1 + \epsilon_2 N_2)}} \quad (91a)$$

and

$$x_2 = \sqrt{\frac{2\epsilon_1 \epsilon_2 N_1 (V_{bi} - V)}{q N_2 (\epsilon_1 N_1 + \epsilon_2 N_2)}}. \quad (91b)$$

► **EXAMPLE 10**

Consider an ideal abrupt heterojunction with a built-in potential of 1.6 V. The impurity concentrations in semiconductor 1 and 2 are 1×10^{16} donors/cm³ and 3×10^{19} acceptors/cm³, and the dielectric constants are 12 and 13, respectively. Find the electrostatic potential and depletion width in each material at thermal equilibrium.

SOLUTION From Eq. 90, the electrostatic potentials of a heterojunction at thermal equilibrium or $V = 0$ are

$$V_{b1} = \frac{13 \times (3 \times 10^{19}) \times 1.6}{12 \times (1 \times 10^{16}) + 13 \times (3 \times 10^{19})} = 1.6 \text{ V}$$

and

$$V_{b2} = \frac{12 \times (1 \times 10^{16}) \times 1.6}{12 \times (1 \times 10^{16}) + 13 \times (3 \times 10^{19})} = 4.9 \times 10^{-4} \text{ V}$$

The depletion widths can be calculated by Eq. 91:

$$x_1 = \sqrt{\frac{2 \times 12 \times 13 \times (8.85 \times 10^{-14}) \times (3 \times 10^{19}) \times 1.6}{(1.6 \times 10^{-19}) \times (1 \times 10^{16}) \times (12 \times 1 \times 10^{16} + 13 \times 3 \times 10^{19})}} = 4.608 \times 10^{-5} \text{ cm},$$

$$x_2 = \sqrt{\frac{2 \times 12 \times 13 \times (8.85 \times 10^{-14}) \times (1 \times 10^{16}) \times 1.6}{(1.6 \times 10^{-19}) \times (3 \times 10^{19}) \times (12 \times 1 \times 10^{16} + 13 \times 3 \times 10^{19})}} = 1.536 \times 10^{-8} \text{ cm}.$$

We see that most of the built-in potential is in the semiconductor with a lower doping concentration. The depletion width there is also much wider. ◀

► SUMMARY

A p - n junction is formed when a p -type and an n -type semiconductor are brought into intimate contact. The p - n junction, in addition to being a device used in many applications, is the basic building block for other semiconductor devices. Therefore, an understanding of junction theory serves as the foundation to understanding other semiconductor devices.

When a p - n junction is formed, there are uncompensated negative ions (N_A^-) on the p -side and uncompensated positive ions (N_D^+) on the n -side. Therefore, a depletion region (i.e., depletion of mobile carriers) is formed at the junction. This region, in turn, creates an electric field. At thermal equilibrium, the drift current due to the electric field is exactly balanced by the diffusion current due to concentration gradients of the mobile carriers on the two sides of the junction. When a positive voltage is applied to the p -side with respect to the n -side, a large current will flow through the junction. However, when a negative voltage is applied, virtually no current flows. This “rectifying” behavior is the most important characteristic of p - n junctions.

The basic equations presented in Chapters 1 and 2 have been used to develop the ideal static and dynamic behaviors of p - n junctions. We derived expressions for the depletion region, the depletion capacitance, and the ideal current-voltage characteristics of p - n junctions. However, practical devices depart from these ideal characteristics because of carrier generation and recombination in the depletion layers, high injection under forward bias, and series-resistance effects. The theory and methods of calculating the effects of these departures from the ideal are discussed in detail. We also considered other factors that influence p - n junctions, such as minority-carrier storage, diffusion capacitance, and transient behavior in high-frequency and switching applications.

A limiting factor in the operation of p - n junctions is junction breakdown—especially that due to avalanche multiplication. When a sufficiently large reverse voltage is applied to a p - n junction, the junction breaks down and conducts a very large current. Therefore, the breakdown voltage imposes an upper limit on the reverse bias for p - n junctions. We derived equations for the breakdown condition of the p - n junction and showed the effect of device geometry and doping on the breakdown voltage.

A related device is the heterojunction formed between two dissimilar semiconductors. We obtained expressions for its electrostatic potentials and depletion widths. These expressions are simplified to that for a conventional p - n junction when these two semiconductors become identical.

► REFERENCES

1. W. Shockley, *Electrons and Holes in Semiconductors*, Van Nostrand, Princeton, NJ, 1950.
2. A. S. Grove, *Physics and Technology of Semiconductor Devices*, Wiley, New York, 1967.
3. R. H. Kingston, “Switching Time in Junction Diodes and Junction Transistors,” *Proc. IRE*, **42**, 829 (1954).
4. J. L. Moll, *Physics of Semiconductors*, McGraw-Hill, New York, 1964.
5. S. M. Sze and G. Gibbons, “Avalanche Breakdown Voltages of Abrupt and Linearly Graded p - n Junctions in Ge, Si, GaAs and GaP,” *Appl. Phys. Lett.*, **8**, 111 (1966).

6. S. K. Ghandhi, *Semiconductor Power Devices*, Wiley, New York, 1977.
7. S. M. Sze and G. Gibbons, "Effect of Junction Curvature on Breakdown Voltages in Semiconductors," *Solid State Electron.*, **9**, 831 (1966).
8. H. Kroemer, "Critique of Two Recent Theories of Heterojunction Lineups," *IEEE Electron Device Lett.*, **EDL-4**, 259 (1983).

► **PROBLEMS (* DENOTES DIFFICULT PROBLEMS)**

FOR SECTION 3.2 DEPLETION REGION

- *1. A diffused silicon *p-n* junction has a linearly graded junction on the *p*-side with $a = 10^{19} \text{ cm}^{-4}$ and a uniform doping of $3 \times 10^{14} \text{ cm}^{-3}$ on the *n*-side. If the depletion layer width of the *p*-side is $0.8 \mu\text{m}$ at zero bias, find the total depletion layer width, built-in potential, and maximum field at zero bias.
- *2. Sketch the potential distribution in the Si *p-n* junction in Prob. 1.
3. For an ideal silicon *p-n* abrupt junction with $N_A = 10^{17} \text{ cm}^{-3}$ and $N_D = 10^{15} \text{ cm}^{-3}$, (a) calculate V_{bi} at 250, 300, 350, 400, 450, and 500 K and plot V_{bi} versus T ; (b) comment on your result in terms of the energy-band diagram; and (c) find the depletion layer width and the maximum field at zero bias for $T = 300 \text{ K}$.
4. Determine the *n*-type doping concentration to meet the following specifications for a Si *p-n* junction: $N_A = 10^{18} \text{ cm}^{-3}$, $\mathcal{E}_{\text{max}} = 4 \times 10^5 \text{ V/cm}$ at $V_R = 30 \text{ V}$, $T = 300 \text{ K}$.

FOR SECTION 3.3 DEPLETION CAPACITANCE

- *5. An abrupt *p-n* junction has a doping concentration of 10^{15} , 10^{16} , or 10^{17} cm^{-3} on the lightly doped *n*-side and of 10^{19} cm^{-3} on the heavily doped *p*-side. Obtain a series of curves of $1/C^2$ versus V , where V ranges from -4 V to 0 V in steps of 0.5 V . Comment on the slopes and the interceptions at the voltage axis of these curves.
6. For a silicon linearly graded junction with a impurity gradient of 10^{20} cm^{-4} , calculate the built-in potential and the junction capacitance at a reverse bias of 4 V ($T = 300 \text{ K}$).
7. A one-sided *p⁺-n* Si junction at 300 K is doped with $N_A = 10^{19} \text{ cm}^{-3}$. Design the junction so that $C_j = 0.85 \text{ pF}$ at $V_R = 4.0 \text{ V}$.

FOR SECTION 3.4 CURRENT-VOLTAGE CHARACTERISTICS

8. Assume that the *p-n* junction considered in Prob. 3 contains 10^{15} cm^{-3} generation-recombination centers located 0.02 eV above the intrinsic Fermi level of silicon with $\sigma_n = \sigma_p = 10^{-15} \text{ cm}^2$. If $v_{th} \cong 10^7 \text{ cm/s}$, calculate the generation and recombination current at -0.5 V .
9. Consider a Si *p-n* junction with an *n*-type doping concentration of 10^{16} cm^{-3} and forward biased with $V = 0.8 \text{ V}$ at 300 K . Calculate the minority-carrier hole concentration at the edge of the space charge region.
10. Calculate the applied reverse-bias voltage at which the ideal reverse current in a *p-n* junction diode at $T = 300 \text{ K}$ reaches 95% of its reverse saturation current value.
11. Design the Si *p-n* diode so that $J_n = 25 \text{ A/cm}^2$ and $J_p = 7 \text{ A/cm}^2$ at $V_a = 0.7 \text{ V}$. The remaining parameters are given in Ex. 5.
12. An ideal silicon *p-n* junction has $N_D = 10^{18} \text{ cm}^{-3}$, $N_A = 10^{16} \text{ cm}^{-3}$, $\tau_p = \tau_n = 10^{-6} \text{ s}$, and a device area of $1.2 \times 10^{-5} \text{ cm}^2$. (a) Calculate the theoretical saturation current at 300 K . (b) Calculate the forward and reverse currents at $\pm 0.7 \text{ V}$.

13. In Prob. 12, assume the widths of the two sides of the junction are much greater than the respective minority-carrier diffusion length. Calculate the applied voltage at a forward current of 1 mA at 300 K.
14. A silicon p^+n junction has the following parameters at 300 K: $\tau_p = \tau_n = 10^{-6}$ s, $N_D = 10^{15}$ cm $^{-3}$, $N_A = 10^{19}$ cm $^{-3}$. (a) Plot diffusion current density, J_{gen} , and total current density versus applied reverse voltage. (b) Repeat the plots for $N_D = 10^{17}$ cm $^{-3}$.

FOR SECTION 3.5 CHARGE STORAGE AND TRANSIENT BEHAVIOR

15. For an ideal abrupt silicon p^+n junction with $N_D = 10^{16}$ cm $^{-3}$, find the stored minority carriers per unit area in the neutral n -region when a forward bias of 1 V is applied. The length of neutral region is 1 μ m and the diffusion length of the holes is 5 μ m.

FOR SECTION 3.6 JUNCTION BREAKDOWN

16. For a silicon p^+n one-sided abrupt junction with $N_D = 10^{15}$ cm $^{-3}$, find the depletion layer width at breakdown. If the n -region is reduced to 5 μ m, calculate the breakdown voltage and compare your result with Fig. 27.
17. Design an abrupt Si p^+n junction diode that has a reverse breakdown voltage of 130 V and a forward-bias current of 2.2 mA at $V_a = 0.7$ volt. Assume $\tau_{po} = 10^{-7}$ s.
18. In Fig. 18b, the avalanche breakdown voltage increases with increasing temperature. Give a qualitative argument for the result.
19. If $\alpha_n = \alpha_p = 10^4(\mathcal{E}/4 \times 10^5)^6$ cm $^{-1}$ in gallium arsenide, where \mathcal{E} is in V/cm, find the breakdown voltage of (a) a $p-i-n$ diode with an intrinsic-layer width of 10 μ m and (b) p^+n junction with a doping of 2×10^{16} cm $^{-3}$ for the lightly doped side.
20. Consider that a Si $p-n$ junction at 300 K with a linearly doping profile varies from $N_A = 10^{18}$ cm $^{-3}$ to $N_D = 10^{18}$ cm $^{-3}$ over a distance of 2 μ m. Calculate the breakdown voltage.

FOR SECTION 3.7 HETEROJUNCTION

21. For the ideal heterojunction in Ex. 10, find the electrostatic potential and depletion width in each material for applied voltages of 0.5 V and -5 V.
22. For an n -type GaAs/ p -type Al $_{0.3}$ Ga $_{0.7}$ As heterojunction at room temperature, $\Delta E_C = 0.21$ eV. Find the total depletion width at thermal equilibrium when both sides have impurity concentration of 5×10^{15} cm $^{-3}$. (Hint: the bandgap of Al $_x$ Ga $_{1-x}$ As is given by $E_g(x) = 1.424 + 1.247x$ eV, and the dielectric constant is $12.4 - 3.12x$. Assume N_C and N_V are the same for Al $_x$ Ga $_{1-x}$ As with $0 < x < 0.4$.)

Bipolar Transistors and Related Devices

- ▶ 4.1 TRANSISTOR ACTION
- ▶ 4.2 STATIC CHARACTERISTICS OF BIPOLAR TRANSISTORS
- ▶ 4.3 FREQUENCY RESPONSE AND SWITCHING OF BIPOLAR TRANSISTORS
- ▶ 4.4 NONIDEAL EFFECTS
- ▶ 4.5 HETEROJUNCTION BIPOLAR TRANSISTORS
- ▶ 4.6 THYRISTORS AND RELATED POWER DEVICES
- ▶ SUMMARY

The transistor (a contraction for *transfer resistor*) is a multijunction semiconductor device. Normally, the transistor is integrated with other circuit elements for voltage gain, current gain, or signal-power gain. The bipolar transistor, also called the bipolar junction transistor (BJT), is one of the most important semiconductor devices. It has been used extensively in high-speed circuits, analog circuits, and power applications. Bipolar devices are semiconductor devices in which both electrons and holes participate in the conduction process. This is in contrast to the field-effect devices, discussed in Chapters 5, 6, and 7, in which predominantly only one kind of carrier participates.

The bipolar transistor was invented by a research team at Bell Laboratories¹ in 1947. The device had two metal wires with sharp points making contact with a germanium substrate (see Fig. 3 in Chapter 0). The first transistor was primitive by today's standards, yet it revolutionized the electronics industry and changed our way of life.

In modern bipolar transistors, we have replaced the germanium with silicon and the point contacts with two closely coupled p - n junctions in the form of p - n - p or n - p - n structures. In this chapter, we consider the transistor action of the coupled junctions and derive the static characteristics from the minority carrier distributions in the device. We also discuss the frequency response and switching behavior of the transistor. In addition, we briefly consider the heterojunction bipolar transistor in which one or both p - n junctions are formed between dissimilar semiconductors.

In the final section, a related bipolar device called a thyristor is introduced. The basic thyristor has three closely coupled p - n junctions in the form of a p - n - p - n structure.² The device exhibits bistable characteristics and can be switched between a high-impedance “off” state and a low-impedance “on” state. (The name thyristor is derived from *gas thyatron*, which is a gas-filled tube with similar bistable characteristics.) Because of the two stable state (on and off) and the low power dissipation in these states, thyristors are useful in many applications. We consider the physical operation of the thyristor and a few related switching devices. Furthermore, the various thyristor types and their applications are briefly introduced.

Specifically, we cover the following topics:

- The current gain and modes of operation of bipolar transistors.
- The cutoff frequency and switching time of a bipolar transistor.
- The advantages of heterojunction bipolar transistors.
- The power-handling capability of thyristor and related bipolar devices.

► 4.1 TRANSISTOR ACTION

A perspective view of a discrete $p-n-p$ bipolar transistor is shown in Fig. 1. The transistor is formed by starting with a p -type substrate. An n -type region is thermally diffused through an oxide window into this p -type substrate. A very heavily doped p^+ region is then diffused into the n -type region. Metallic contacts are made to the p^+ - and n -regions through the windows opened in the oxide layer and to the p -region at the bottom. The details of transistor fabrication processes are considered in later chapters.

An idealized, one-dimensional structure of a $p-n-p$ bipolar transistor between the dashed lines in Fig. 1 is shown in Fig. 2a. Normally, the bipolar transistor has three separately doped regions and two $p-n$ junctions. The heavily doped p^+ -region is called the *emitter* (symbol E in the figure). The narrow central n -region, with moderately doped concentration, is called the *base* (symbol B). The width of the base is small compared with the minority-carrier diffusion length. The lightly doped p -region is called the *collector* (symbol C). The doping concentration in each region is assumed to be uniform. Note that the concepts developed for the $p-n$ junction can be applied directly to the transistor.

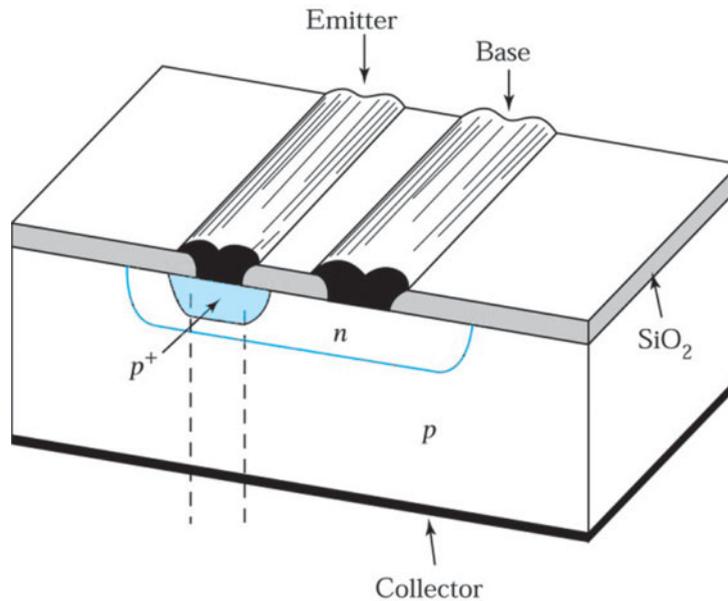


Fig. 1 Perspective view of a silicon $p-n-p$ bipolar transistor.

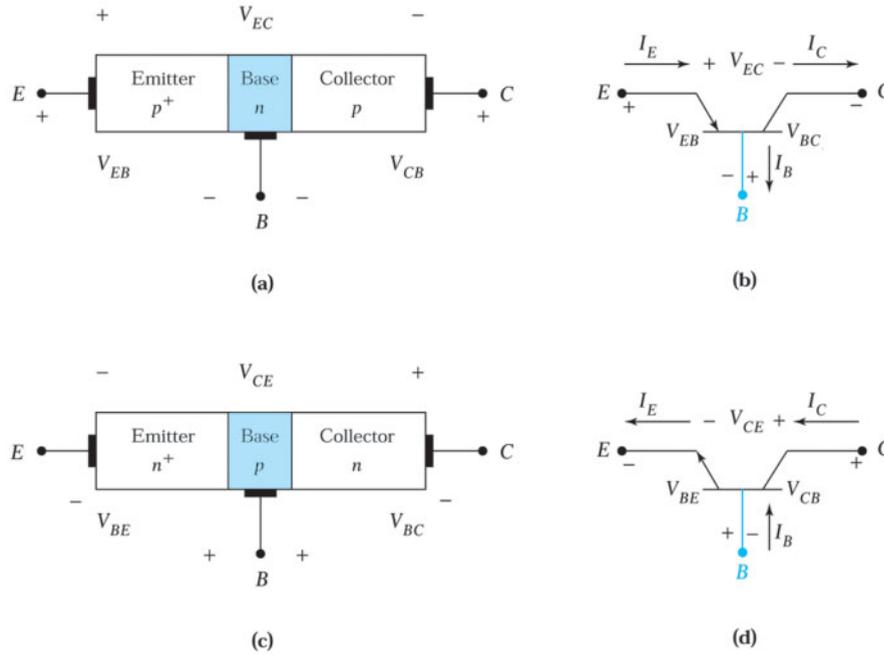


Fig. 2 (a) Idealized one-dimensional schematic of a $p-n-p$ bipolar transistor and (b) its circuit symbol. (c) Idealized one-dimensional schematic of an $n-p-n$ bipolar transistor and (d) its circuit symbol.

Figure 2b shows the circuit symbol for a $p-n-p$ transistor. The current components and voltage polarities are shown in the figure. The arrows of the various currents indicate the direction of current flow under normal operating conditions (also called the *active mode*). The + and – signs are used to define the voltage polarities. We can also denote the voltage polarity by a double subscript on the voltage symbol. In the active mode, the emitter-base junction is forward biased ($V_{EB} > 0$) and the base-collector junction is reverse biased ($V_{CB} < 0$). According to Kirchhoff's circuit laws, there are only two independent currents for this three-terminal device. If two currents are known, the third current can be obtained.

The $n-p-n$ bipolar transistor is the complementary structure to the $p-n-p$ bipolar transistor. The structure and circuit symbol of an ideal $n-p-n$ transistor are shown in Figs. 2c and 2d, respectively. The $n-p-n$ structure can be obtained by interchanging p for n and n for p in the $p-n-p$ transistor. As a result, the current flows and voltage polarities are all reversed. In subsequent sections, we concentrate on the $p-n-p$ type because the direction of minority-carrier (hole) flow is the same as that of current flow. It provides a more intuitive base for understanding the mechanisms of charge transport. Once we understand the $p-n-p$ transistor, we need only reverse the polarities and conduction types to describe the $n-p-n$ transistor.

4.1.1 Operation in the Active Mode

Figure 3a show an idealized $p-n-p$ transistor in thermal equilibrium, that is, when all three leads are connected together or all are grounded. The depletion regions near the two junctions are illustrated by colored areas. Figure 3b shows the impurity densities in the three doped regions, where the emitter is more heavily doped than the collector. However, the base doping is less than the emitter doping but greater than the collector doping. Figure 3c shows the corresponding electric-field profiles in the two depletion regions.

Figure 3d illustrates the energy band diagram, which is a simple extension of the thermal-equilibrium situation for the $p-n$ junction applied to a pair of closely coupled p^+-n and $n-p$ junctions. The results obtained for the $p-n$ junction in Chapter 3 are equally applicable to the emitter-base and base-collector junctions. At thermal equilibrium there is no net current flow, and hence the Fermi level is a constant in the regions.

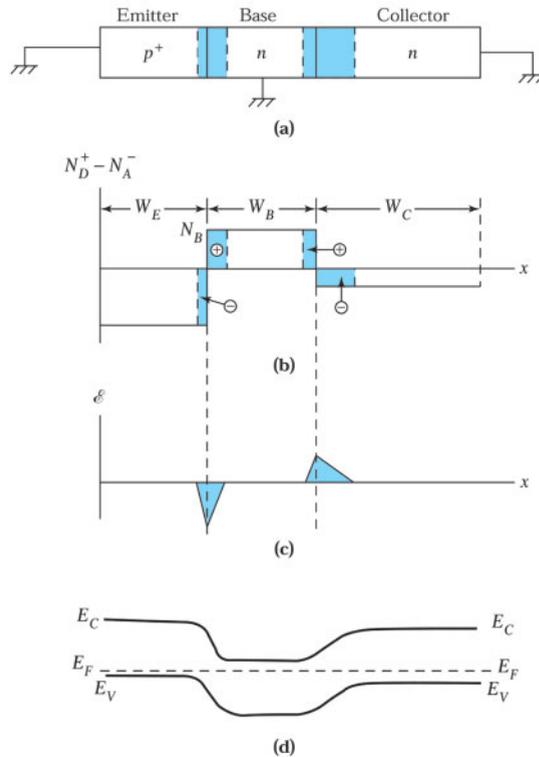


Fig. 3 (a) A p - n - p transistor with all leads grounded (at thermal equilibrium). (b) Doping profile of a transistor with abrupt impurity distributions. (c) Electric-field profile. (d) Energy band diagram at thermal equilibrium.

Figure 4 shows the corresponding cases when the transistor in Fig. 3 is biased in the active mode. Figure 4a is a schematic of the transistor connected as an amplifier with the *common-base configuration*, that is, the base lead is common to the input and output circuits.³ Figures 4b and 4c show the charge densities and the electric fields, respectively, under biasing conditions. Note that the depletion layer width of the emitter-base junction is narrower and the collector-base junction is wider than in the equilibrium case in Fig. 3.

Figure 4d shows the corresponding energy band diagram under the active mode. Since the emitter-base junction is forward biased, holes are injected (or emitted) from the p^+ emitter into the base and electrons are injected from the n base into the emitter. Under the ideal-diode condition, there is no generation-recombination current in the depletion region; these two current components constitute the total emitter current. The collector-base junction is reverse biased and a small reverse saturation current will flow across the junction. However, if the base width is sufficiently small, the holes injected from the emitter can diffuse through the base to reach the base-collector depletion edge and then “float up” into the collector (recall the “bubble analogy”). This transport mechanism gives rise to the terminology of *emitter*, which emits or injects carriers, and *collector*, which collects these carriers injected from a nearby junction. If most of the injected holes can reach the collector without recombining with electrons in the base region, the collector hole current will be very close to the emitter hole current.

Therefore, carriers injected from a nearby emitter junction can result in a large current flow in a reverse-biased collector junction. This is the *transistor action*, and it can be realized only when the two junctions are physically close enough to interact in the manner described. The two junctions are called the *interacting p - n junctions*. If, on the other hand, the two junctions are so far apart that all the injected holes are recombined in the base before reaching the base-collector junction, then the transistor action is lost and the p - n - p structure becomes merely two diodes connected back to back.

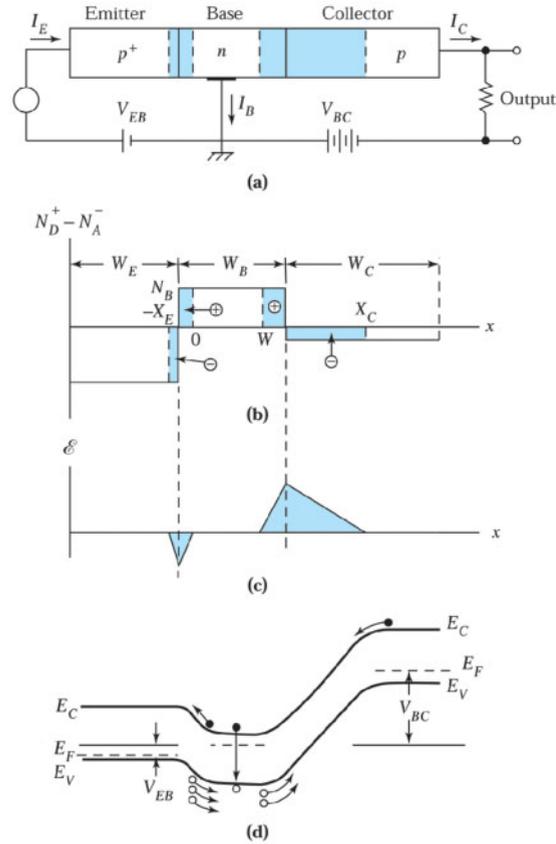


Fig. 4 (a) The transistor shown in Fig. 3 under the active mode of operation.³ (b) Doping profiles and depletion regions under biasing conditions. (c) Electric-field profile. (d) Energy band diagram.

4.1.2 Current Gain

Figure 5 shows the various current components in an ideal p - n - p transistor biased in the active mode. Note that we assume that there are no generation-recombination currents in the depletion regions. The holes injected from the emitter constitute the current I_{Ep} , which is the largest current component in a well-designed transistor. Most of the injected holes will reach the collector junction and give rise to the current I_{Cp} . There are three base current components, labeled I_{BB} , I_{En} , and I_{Cn} . I_{BB} corresponds to electrons that must be supplied by the base to replace electrons recombined with the injected holes (i.e., $I_{BB} = I_{Ep} - I_{Cp}$). I_{En} corresponds to the current arising from electrons being injected from the base to the emitter. However, I_{En} is not desirable, as shown later. It can be minimized by using heavier emitter doping (Section 4.2) or a heterojunction (Section 4.5). I_{Cn} corresponds to thermally generated electrons that are near the base-collector junction edge and drift from the collector to the base. As indicated in the figure, the direction of the electron current is opposite the direction of the electron flow.

We can now express the terminal currents in terms of the various current components described above:

$$I_E = I_{Ep} + I_{En}, \quad (1)$$

$$I_C = I_{Cp} + I_{Cn}, \quad (2)$$

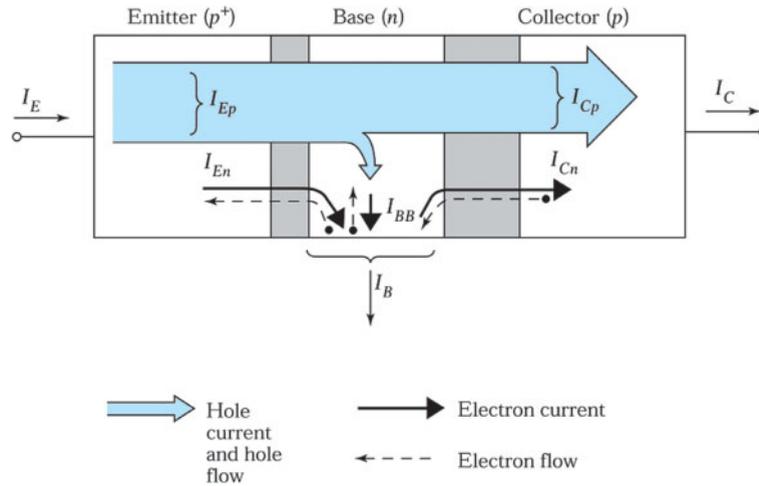


Fig. 5 Various current components in a p - n - p transistor under active mode of operation. The electron flow is in the opposite direction to the electron current.

$$I_B = I_E - I_C = I_{En} + (I_{Ep} - I_{Cp}) - I_{Cn}. \quad (3)$$

An important parameter in the characterization of bipolar transistors is the *common-base current gain* α_0 . This quantity is defined by

$$\alpha_0 \equiv \frac{I_{Cp}}{I_E}. \quad (4)$$

Substituting Eq. 1 into Eq. 4 yields

$$\alpha_0 = \frac{I_{Cp}}{I_{Ep} + I_{En}} = \left(\frac{I_{Ep}}{I_{Ep} + I_{En}} \right) \left(\frac{I_{Cp}}{I_{Ep}} \right). \quad (5)$$

The first term on the right-hand side is called the *emitter efficiency* γ , which is a measure of the injected hole current compared with the total emitter current:

$$\gamma \equiv \frac{I_{Ep}}{I_E} = \frac{I_{Ep}}{I_{Ep} + I_{En}}. \quad (6)$$

The second term is called the *base transport factor* α_T , which is the ratio of the hole current reaching the collector to the hole current injected from the emitter:

$$\alpha_T \equiv \frac{I_{Cp}}{I_{Ep}}. \quad (7)$$

Therefore, Eq. 5 becomes

$$\alpha_0 = \gamma \alpha_T. \quad (8)$$

For a well-designed transistor, because I_{En} is small compared with I_{Ep} and I_{Cp} is close to I_{Ep} , both γ and α_T approach unity. Therefore, α_0 is close to 1.

We can express the collector current in terms of α_0 . The collector current can be described by substituting Eqs. 6 and 7 into Eq. 2:

$$I_C = I_{Cp} + I_{Cn} = \alpha_T I_{Ep} + I_{Cn} = \gamma \alpha_T \left(\frac{I_{Ep}}{\gamma} \right) + I_{Cn} = \alpha_0 I_E + I_{Cn}, \quad (9)$$

where I_{Cn} corresponds to the collector-base current flowing with the emitter open-circuited ($I_E = 0$). We designate I_{Cn} as I_{CBO} , where the first two subscripts (CB) refer to the two terminals between which the current (or voltage) is measured and the third subscript (O) refers to the state of the third terminal with respect to the second. In the present case, I_{CBO} designates the leakage current between the collector and the base with the emitter-base junction open. The collector current for the common-base configuration is then given by

$$\boxed{I_C = \alpha_0 I_E + I_{CBO}} \quad (10)$$

► EXAMPLE 1

For an ideal $p-n-p$ transistor, the current components are given by $I_{Ep} = 3$ mA, $I_{En} = 0.01$ mA, $I_{Cp} = 2.99$ mA, and $I_{Cn} = 0.001$ mA. Determine (a) the emitter efficiency γ , (b) the base transport factor α_T , (c) the common-base current gain α_0 , and (d) I_{CBO} .

SOLUTION

(a) Using Eq. 6, the emitter efficiency is

$$\gamma = \frac{I_{Ep}}{I_{Ep} + I_{En}} = \frac{3}{3 + 0.01} = 0.9967.$$

(b) The base transport factor can be obtained from Eq. 7:

$$\alpha_T = \frac{I_{Cp}}{I_{Ep}} = \frac{2.99}{3} = 0.9967.$$

(c) The common-base current gain is given by Eq. 8:

$$\alpha_0 = \gamma \alpha_T = 0.9967 \times 0.9967 = 0.9933.$$

(d)

$$I_E = I_{Ep} + I_{En} = 3 + 0.01 = 3.01 \text{ mA}$$

$$I_C = I_{Cp} + I_{Cn} = 2.99 + 0.001 = 2.991 \text{ mA}$$

Using Eq. 10, we find

$$I_{CBO} = I_C - \alpha_0 I_E = 2.991 - 0.9933 \times 3.01 = 0.001 \text{ mA} . \quad \blacktriangleleft$$

► 4.2 STATIC CHARACTERISTICS OF BIPOLAR TRANSISTORS

In this section, we study the static current-voltage characteristics for an ideal transistor and derive equations for the terminal currents. The current equations are based on the minority-carrier concentration in each region and therefore are described by semiconductor parameters such as doping and minority-carrier lifetime.

4.2.1 Carrier Distribution in Each Region

To derive the current-voltage expression for an ideal transistor, we assume the following:

1. The device has uniform doping in each region.
2. The hole drift current in the base region as well as the collector saturation current is negligible.
3. There is low-level injection.
4. There are no generation-combination currents in the depletion regions.
5. There are no series resistances in the device.

Basically, we assume that holes are injected from the emitter into the base under the forward-biased condition. These holes then diffuse across the base region and reach the collector junction. Once we determine the minority-carrier distribution (i.e., holes in the n -type base region), we can obtain the current from the minority-carrier gradient.

Base Region

Figure 4c shows the electric-field distributions across the junction depletion regions. The minority-carrier distribution in the neutral base region can be described by the field-free, steady-state continuity equation

$$D_p \left(\frac{d^2 p_n}{dx^2} \right) - \frac{p_n - p_{no}}{\tau_p} = 0, \quad (11)$$

where D_p and τ_p are the diffusion constant and the lifetime of minority carriers, respectively. The general solution of Eq. 11 is

$$p_n(x) = p_{no} + C_1 e^{x/L_p} + C_2 e^{-x/L_p}, \quad (12)$$

where $L_p = \sqrt{D_p \tau_p}$ is the diffusion length of holes. The constants C_1 and C_2 can be determined by the boundary conditions for the active mode:

$$p_n(0) = p_{no} e^{qV_{EB}/kT} \quad (13a)$$

and

$$p_n(W) = 0, \quad (13b)$$

where p_{no} is the equilibrium minority-carrier concentration in the base, given by $p_{no} = n_i^2 / N_B$, and N_B denotes the uniform donor concentration in the base. The first boundary condition (Eq. 13a) states that under forward bias, the minority-carrier concentration at the edge of the emitter-base depletion region ($x = 0$) is increased above the equilibrium value by the exponential factor $e^{qV_{EB}/kT}$. The second boundary condition (Eq. 13b) states that under reverse bias, the minority carrier concentration at the edge of the base-collector depletion region ($x = W$) is zero.

Substituting Eq. 13 into the general solution expressed in Eq. 12 yields

$$p_n(x) = p_{no} \left(e^{qV_{EB}/kT} - 1 \right) \left[\frac{\sinh\left(\frac{W-x}{L_p}\right)}{\sinh\left(\frac{W}{L_p}\right)} \right] + p_{no} \left[1 - \frac{\sinh\left(\frac{x}{L_p}\right)}{\sinh\left(\frac{W}{L_p}\right)} \right]. \quad (14)$$

The sinh function, $\sinh(\Lambda)$, can be approximately expressed by Λ when $\Lambda \ll 1$. For example, when $\Lambda < 0.3$, the difference between $\sinh(\Lambda)$ and Λ is less than 1.5 percent. Therefore, when $W/L_p \ll 1$, the distribution equation can be simplified as

$$p_n(x) = p_{no} e^{qV_{EB}/kT} \left(1 - \frac{x}{W}\right) = p_n(0) \left(1 - \frac{x}{W}\right). \quad (15)$$

The distribution approaches a straight line. The approximation is reasonable because the width of the base region is designed to be much smaller than the diffusion length of the minority carrier. Figure 6 shows a linear minority-carrier distribution in a typical transistor operated under active mode. Note that assuming linear minority-carrier distribution can simplify the derivation of current-voltage characteristics. Therefore, we use the assumption hereafter to derive equations for the current-voltage characteristics.

Emitter and Collector Regions

The minority-carrier distributions in the emitter and collector can be obtained in a manner similar to the one used to obtain the distributions for the base region. In Fig. 6, the boundary conditions in the neutral emitter and collector regions are

$$n_E(x = -x_E) = n_{EO} e^{qV_{EB}/kT} \quad (16)$$

and

$$n_C(x = x_C) = n_{CO} e^{-qV_{CB}/kT} = 0, \quad (17)$$

where n_{EO} and n_{CO} are the equilibrium electron concentrations in the emitter and collector, respectively. We assume that the emitter depth and the collector depth are much larger than their corresponding diffusion lengths L_E and L_C , respectively. Substituting these boundary conditions into expressions similar to Eq. 12 yields

$$n_E(x) = n_{EO} + n_{EO} \left(e^{qV_{EB}/kT} - 1 \right) e^{\frac{x+x_E}{L_E}} \quad x \leq -x_E, \quad (18)$$

$$n_C(x) = n_{CO} - n_{CO} e^{\frac{x-x_C}{L_C}} \quad x \geq x_C. \quad (19)$$

4.2.2 Ideal Transistor Currents for Active Mode Operation

Once the minority-carrier distributions are known, the various current components shown in Fig. 6 can be calculated. The hole current I_{Ep} , injected from the emitter at $x = 0$, is proportional to the gradient of the minority carrier concentration. For $W/L_p \ll 1$, the hole current I_{Ep} can be expressed by using Eq. 15:

$$I_{Ep} = A \left(-qD_p \frac{dp_n}{dx} \Big|_{x=0} \right) \cong \frac{qAD_p p_{no}}{W} e^{qV_{EB}/kT}. \quad (20)$$

Similarly, the hole current collected by the collector at $x = W$ is

$$I_{Cp} = A \left(-qD_p \frac{dp_n}{dx} \Big|_{x=W} \right) \cong \frac{qAD_p p_{no}}{W} e^{qV_{EB}/kT}. \quad (21)$$

Note that I_{Ep} is equal to I_{Cp} when $W/L_p \ll 1$. The electron current I_{En} , which is due to electron flow from the base to the emitter, and I_{Cn} , which is due to electron flow from the collector to the base, are

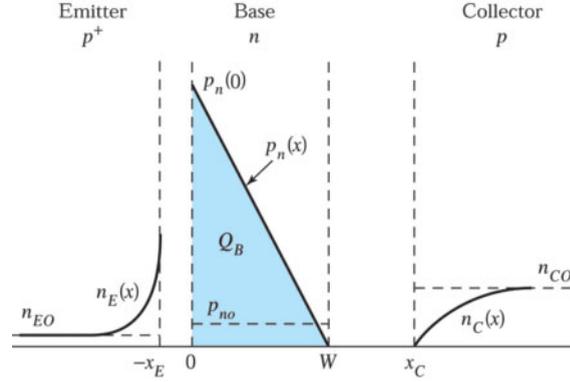


Fig. 6 Minority carrier distribution in various regions of a p - n - p transistor under the active mode of operation.

$$I_{En} = A \left(-qD_E \frac{dn_E}{dx} \Big|_{x=-x_E} \right) = \frac{qAD_E n_{EO}}{L_E} (e^{qV_{EB}/kT} - 1), \quad (22)$$

$$I_{Cn} = A \left(-qD_C \frac{dn_C}{dx} \Big|_{x=x_C} \right) = \frac{qAD_C n_{CO}}{L_C}, \quad (23)$$

where D_E and D_C are the diffusion constants in the emitter and collector, respectively.

The terminal currents can now be obtained from these equations. The emitter current is the sum of Eqs. 20 and 22:

$$I_E = a_{11} (e^{qV_{EB}/kT} - 1) + a_{12} \quad (24)$$

where

$$a_{11} \equiv qA \left(\frac{D_p p_{no}}{W} + \frac{D_E n_{EO}}{L_E} \right), \quad (25)$$

$$a_{12} \equiv \frac{qAD_p p_{no}}{W}. \quad (26)$$

The collector current is the sum of Eqs. 21 and 23:

$$I_C = a_{21} (e^{qV_{EB}/kT} - 1) + a_{22}, \quad (27)$$

where

$$a_{21} \equiv \frac{qAD_p p_{no}}{W}, \quad (28)$$

$$a_{22} \equiv qA \left(\frac{D_p p_{no}}{W} + \frac{D_C n_{CO}}{L_C} \right). \quad (29)$$

Note that $a_{12} = a_{21}$. The base current for the ideal transistor is the difference between the emitter current (I_E) and the collector current (I_C). Therefore, the base current can be obtained by subtracting Eq. 27 from Eq. 24:

$$I_B = (a_{11} - a_{21})(e^{qV_{EB}/kT} - 1) + (a_{12} - a_{22}). \quad (30)$$

From these discussions, we see that the currents in the three terminals of a transistor are mainly determined by the minority carrier distribution in the base region. Once we derive the current components, the common-base current gain α_0 can be obtained by using Eqs. 6 through 8.

► EXAMPLE 2

An ideal p^+n-p transistor has impurity concentrations of 10^{19} , 10^{17} , and $5 \times 10^{15} \text{ cm}^{-3}$ in the emitter, base, and collector regions, respectively; the corresponding lifetimes are 10^{-8} , 10^{-7} , and 10^{-6} s. Assume that an effective cross section area A is 0.05 mm^2 and the emitter-base junction is forward-biased to 0.6 V . Find the common-base current gain of the transistor. Note that the other device parameters are $D_E = 1 \text{ cm}^2/\text{s}$, $D_p = 10 \text{ cm}^2/\text{s}$, $D_C = 2 \text{ cm}^2/\text{s}$, and $W = 0.5 \text{ }\mu\text{m}$.

SOLUTION In the base region,

$$L_p = \sqrt{D_p \tau_p} = \sqrt{10 \times 10^{-7}} = 10^{-3} \text{ cm},$$

$$p_{no} = n_i^2 / N_B = (9.65 \times 10^9)^2 / 10^{17} = 9.31 \times 10^2 \text{ cm}^{-3}.$$

Similarly, in the emitter region, $L_E = \sqrt{D_E \tau_E} = 10^{-4} \text{ cm}$ and $n_{EO} = n_i^2 / N_E = 9.31 \text{ cm}^{-3}$. Since $W/L_p = 0.05 \ll 1$, the current components are given by

$$I_{Ep} = \frac{1.6 \times 10^{-19} \times 5 \times 10^{-4} \times 10 \times 9.31 \times 10^2}{0.5 \times 10^{-4}} \times e^{0.6/0.0259} \text{ A} = 1.7137 \times 10^{-4} \text{ A},$$

$$I_{Cp} = 1.7137 \times 10^{-4} \text{ A},$$

$$I_{En} = \frac{1.6 \times 10^{-19} \times 5 \times 10^{-4} \times 1 \times 9.31}{10^{-4}} (e^{0.6/0.0259} - 1) = 8.5687 \times 10^{-8} \text{ A}.$$

Therefore, the common-base current gain α_0 is

$$\alpha_0 = \frac{I_{Cp}}{I_{Ep} + I_{En}} = \frac{1.7137 \times 10^{-4}}{1.7137 \times 10^{-4} + 8.5687 \times 10^{-8}} = 0.9995.$$

For the case of $W/L_p \ll 1$, we can simplify the emitter efficiency from Eqs. 20 and 22:

$$\gamma \equiv \frac{I_{Ep}}{I_{Ep} + I_{En}} \cong \frac{\frac{D_p p_{no}}{W}}{\frac{D_p p_{no}}{W} + \frac{D_E n_{EO}}{L_E}} = \frac{1}{1 + \frac{D_E n_{EO} W}{D_p p_{no} L_E}} \quad (31)$$

or

$$\gamma = \frac{1}{1 + \frac{D_E}{D_p} \cdot \frac{N_B}{N_E} \cdot \frac{W}{L_E}}, \quad (31a)$$

where $N_B (= n_i^2 / p_{no})$ is the impurity doping in the base and $N_E (= n_i^2 / n_{EO})$ is the impurity doping in the emitter. This equation shows that to improve γ , we should decrease the ratio N_B/N_E ; that is, there should be much heavier doping in the emitter than in the base. This is the reason we use p^+ -doping in the emitter.

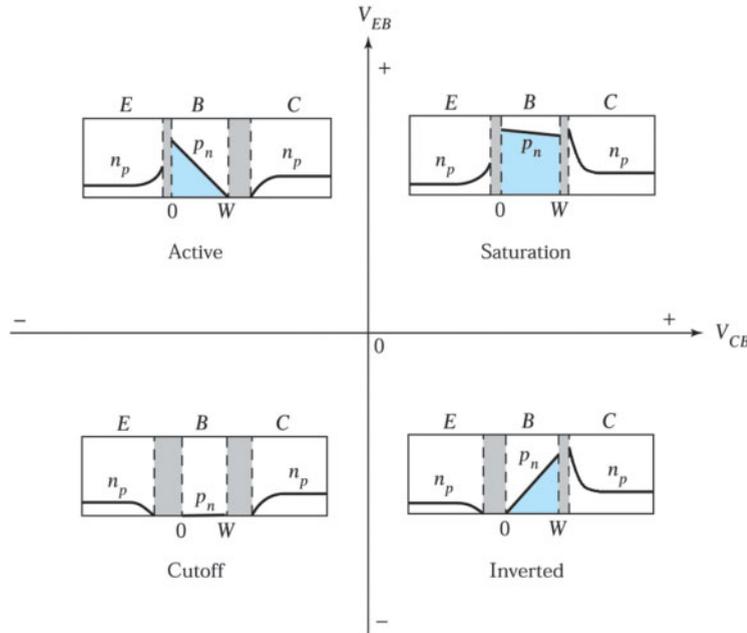


Fig. 7 Junction polarities and minority carrier distributions of a p - n - p transistor under four modes of operation.

4.2.3 Modes of Operation

A bipolar transistor has four modes of operation, depending on the voltage polarities on the emitter-base junction and the collector-base junction. Figure 7 shows the V_{EB} and V_{CB} voltages for the four modes of operations of a p - n - p transistor. The corresponding minority carrier distributions are also shown. So far in this chapter we have considered the *active mode* of transistor operation. In the active mode, the emitter-base junction is forward biased and the base-collector junction is reverse biased.

In the *saturation mode*, both junctions are forward biased, leading to the nonzero minority-carrier distribution at the edge of each depletion region. Therefore, the boundary condition at $x = W$ becomes $p_n(W) = p_{n0} e^{qV_{CB}/kT}$ instead of that given by Eq. 13b. The saturation mode corresponds to small biasing voltage and large output current, that is, the transistor is in a conducting state and acts as a closed (or on) switch.

In the *cutoff mode*, both junctions are reverse biased. The boundary conditions of Eq. 13 become $p_n(0) = p_n(W) = 0$. The cutoff mode corresponds to the open (or off) state of the transistor as a switch.

The fourth mode of operation is the *inverted mode*, called the inverted active mode. In this mode, the emitter-base junction is reverse biased and the collector-base junction is forward biased. The inverted mode corresponds to the case where the collector acts as the emitter and the emitter acts as a collector. In this condition, the device is used backward. However, the current gain for the inverted mode is generally lower than that for the active mode. It is because of poor “emitter efficiency” resulting from low collector doping with respect to the base doping (Eq. 31).

The current-voltage relationships for the various modes of operation can be obtained by following the same procedures used for the active mode, with an appropriate change in the boundary conditions as in Eq. 13. The general expressions applicable to all modes of operations are

$$I_E = a_{11}(e^{qV_{EB}/kT} - 1) - a_{12}(e^{qV_{CB}/kT} - 1) \quad (32a)$$

and

$$I_C = a_{21}(e^{qV_{EB}/kT} - 1) - a_{22}(e^{qV_{CB}/kT} - 1) \quad (32b)$$

where the coefficients a_{11} , a_{12} , a_{21} , and a_{22} are given by Eqs. 25, 26, 28, and 29, respectively. Note that in Eqs. 32a and 32b the biasing voltages for the junctions can be positive or negative depending on the mode of operation.

4.2.4 Current-Voltage Characteristics of Common-Base and Common-Emitter Configurations

Using Eq. 32, we can obtain the current-voltage characteristics for a transistor in a common-base configuration. Note that in this configuration, V_{EB} and V_{BC} are the input and output voltages and I_E and I_C are the input and output currents, respectively.

However, in circuit applications the common-emitter configuration is the one most often used, where the emitter lead is common to the input and output circuits. The general expressions for the currents, shown in Eq. 32, are also applicable to the common-emitter configuration. In this case, to generate the current-voltage characteristics, V_{EB} and I_B are the input parameters and V_{EC} and I_C are the output parameters.

Common-Base Configuration

Figure 8a shows the common-base configuration of a p - n - p transistor. Figure 8b shows the measured results of output current-voltage characteristics for the common-base configuration. The various modes of operation are indicated on the figure. Note that the measured results of the output current-voltage characteristics for the common-base collector current are practically equal to the emitter current (i.e., $\alpha_0 \cong 1$) and virtually independent of V_{BC} . This is in close agreement with the ideal transistor behavior given by Eqs. 10 and 27. The collector current remains practically constant, even down to zero volts for V_{BC} , where the holes are still extracted by the collector. This is indicated by the hole distributions shown in Fig. 9a. Since the hole gradient at $x = W$ changes only slightly from $V_{BC} > 0$ to $V_{BC} = 0$, the collector current remains essentially the same over the entire active mode of operation. To reduce the collector current to zero, we have to apply a small forward bias, about 1 V for silicon, to the base-collector junction (in the saturation mode), as shown in Fig. 9b. The forward bias will sufficiently increase the hole density at $x = W$ to make it equal to that of the emitter at $x = 0$ (see the horizontal line in Fig. 9b). Therefore, the hole gradient at $x = W$ as well as the collector current will be reduced to zero.

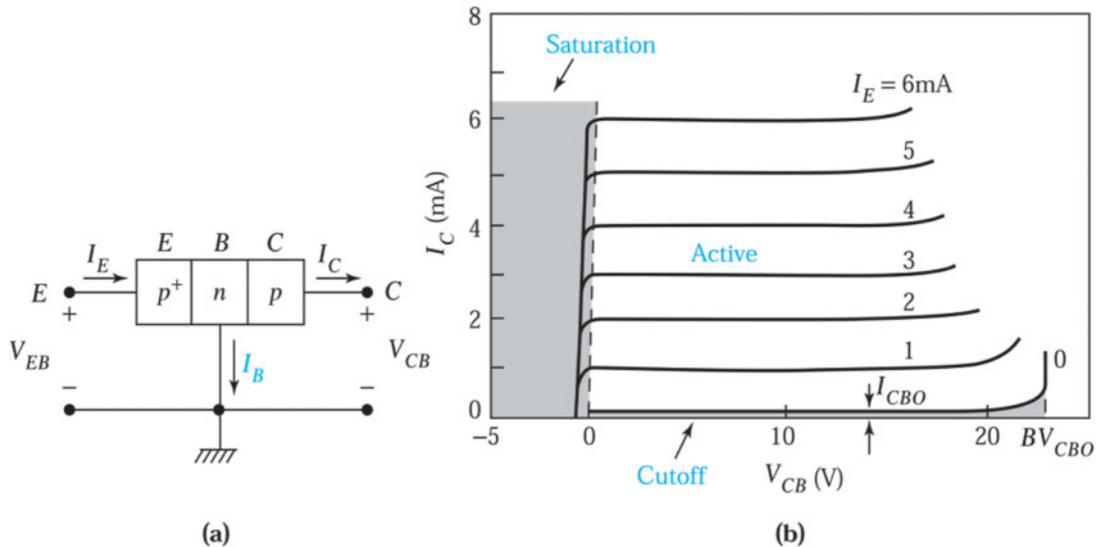


Fig. 8 (a) Common-base configuration of a p - n - p transistor. (b) Its output current-voltage characteristics.

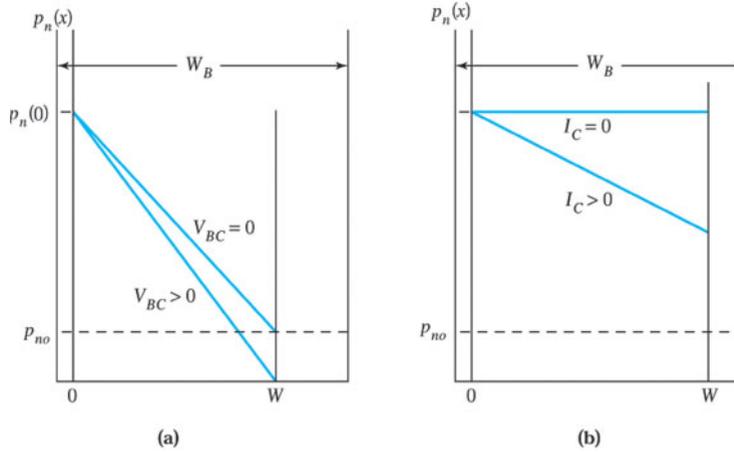


Fig. 9 Minority carrier distributions in the base region of a $p-n-p$ transistor. (a) Active mode for $V_{BC} = 0$ and $V_{BC} > 0$. (b) Saturation mode with both junctions forward biased.

Common-Emitter Configuration

Figure 10a shows the common-emitter configuration for a $p-n-p$ transistor. The collector current for the common-emitter configuration can be obtained by substituting Eq. 3 into Eq. 10:

$$I_C = \alpha_0(I_B + I_C) + I_{CBO}. \quad (33)$$

Solving for I_C , we obtain

$$I_C = \frac{\alpha_0}{1 - \alpha_0} I_B + \frac{I_{CBO}}{1 - \alpha_0}. \quad (34)$$

We now designate β_0 as the *common-emitter current gain*, which is the incremental change of I_C with respect to an incremental change of I_B . From Eq. 34, we obtain

$$\beta_0 \equiv \frac{\Delta I_C}{\Delta I_B} = \frac{\alpha_0}{1 - \alpha_0}. \quad (35)$$

We can also designate I_{CEO} as

$$I_{CEO} \equiv \frac{I_{CBO}}{1 - \alpha_0}. \quad (36)$$

This current corresponds to the collector-emitter leakage current for $I_B = 0$. Equation 34 becomes

$$I_C = \beta_0 I_B + I_{CEO}. \quad (37)$$

Because the value of α_0 is generally close to unity, β_0 is much larger than 1. For example, if $\alpha_0 = 0.99$, β_0 is 99 and if α_0 is 0.998, β_0 is 499. Therefore, a small change in the base current can give rise to a much larger change in the collector current. Figure 10b shows the measured results of output current-voltage characteristics with various input base currents. Note that the figure shows nonzero collector-emitter leakage current I_{CEO} when $I_B = 0$.

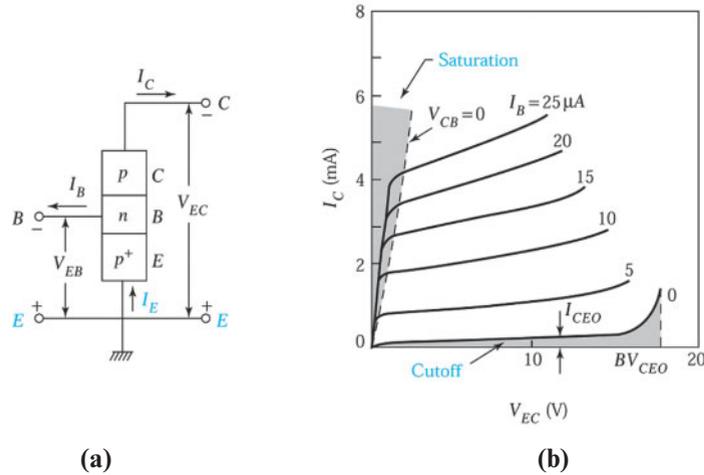


Fig. 10 (a) Common-emitter configuration of a p - n - p transistor. (b) Its output current-voltage characteristics.

► EXAMPLE 3

Referring to Ex. 1, find the common-emitter current gain β_0 . Express I_{CEO} in terms of β_0 and I_{CBO} and find the value of I_{CEO} .

SOLUTION The common-base current gain α_0 in Example 1 is 0.9933. Hence, we can obtain β_0 by

$$\beta_0 = \frac{0.9933}{1 - 0.9933} = 148.3.$$

Equation 36 can be expressed by

$$\begin{aligned} I_{CEO} &= \left(\frac{\alpha_0}{1 - \alpha_0} + 1 \right) I_{CBO} \\ &= (\beta_0 + 1) I_{CBO}. \end{aligned}$$

Therefore,

$$I_{CEO} = (148.3 + 1) \times 1 \times 10^{-6} = 1.49 \times 10^{-4} \text{ A}$$

In an ideal transistor with the common-emitter configuration, the collector current for a given I_B is expected to be independent of V_{EC} for $V_{EC} > 0$. This is true when we assume that the neutral base width (W) is constant. However, since the width of the space charge region extending into the base region varies with the base-collector voltage, the base width is a function of the base-collector voltage. The collector current, therefore, is dependent on V_{EC} . As the base-collector reverse-bias voltage increases, the base width will be reduced as shown in Fig. 11a. The reduced base width causes the gradient in the minority-carrier concentration to increase, which causes an increase in the diffusion current. As a result, β_0 will be increased. Figure 11b shows pronounced slopes and I_C increasing with increasing V_{EC} . This deviation is known as the *Early effect*⁴ or the *base width modulation*. By extrapolating the collector currents and intersecting the V_{EC} axis, we can obtain the voltage V_A , which is called the *Early voltage*.

► 4.3 FREQUENCY RESPONSE AND SWITCHING OF BIPOLAR TRANSISTORS

In Section 4.2 we discussed four possible modes of operation that depend on the biasing conditions of the emitter-base and collector-base junctions. Generally, in analog or linear circuits the transistors are operated in the active mode only. However, in digital circuits all four modes of operation may be involved. In this section we consider the frequency response and switching characteristics of bipolar transistors.

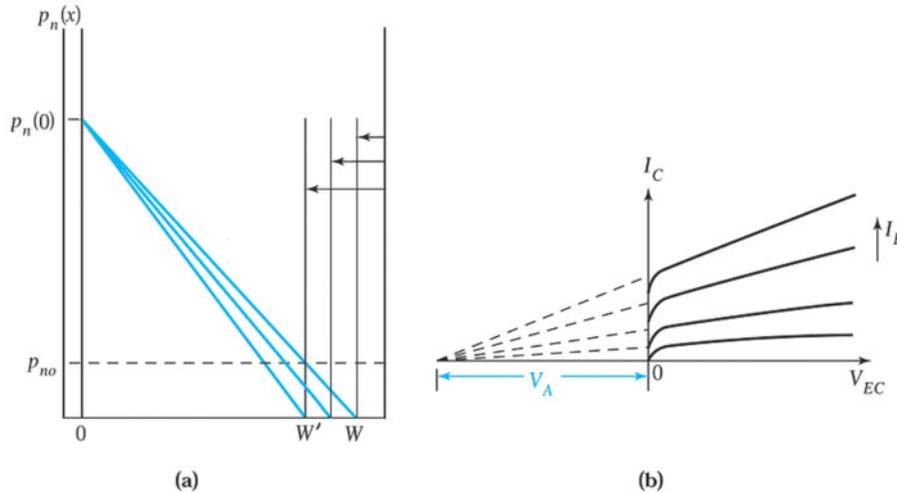


Fig. 11 Schematic diagram of (a) the Early effect and (b) Early voltage V_A . The collector currents for different base currents meet at $-V_A$.

4.3.1 Frequency Response

High-Frequency Equivalent Circuit

In previous discussions, we were concerned with the static [or direct current (dc)] characteristics of the bipolar transistor. We now study its alternating current (ac) characteristics when a small-signal voltage or current is superimposed upon the dc values. The term small-signal means that the peak values of the ac signal current and voltage are smaller than the dc values. Consider an amplifier circuit shown in Fig. 12a, where the transistor is connected in a common-emitter configuration. For a given dc input voltage V_{EB} , a dc base current I_B and dc collector current I_C flow in the transistor. These currents correspond to the operating point shown in Fig. 12b. The load line, determined by the applied voltage V_{CC} and the load resistance R_L , intercepts the V_{EC} axis at V_{CC} and has a slope of $(-1/R_L)$. When a small ac signal is superimposed on the input voltage, the base current I_B will vary as a function of time, as illustrated in Fig. 12b. This variation, in turn, brings about a corresponding variation in the output current i_C , which is β_0 times larger than the input current variation. As a result, the transistor amplifies the input signal.

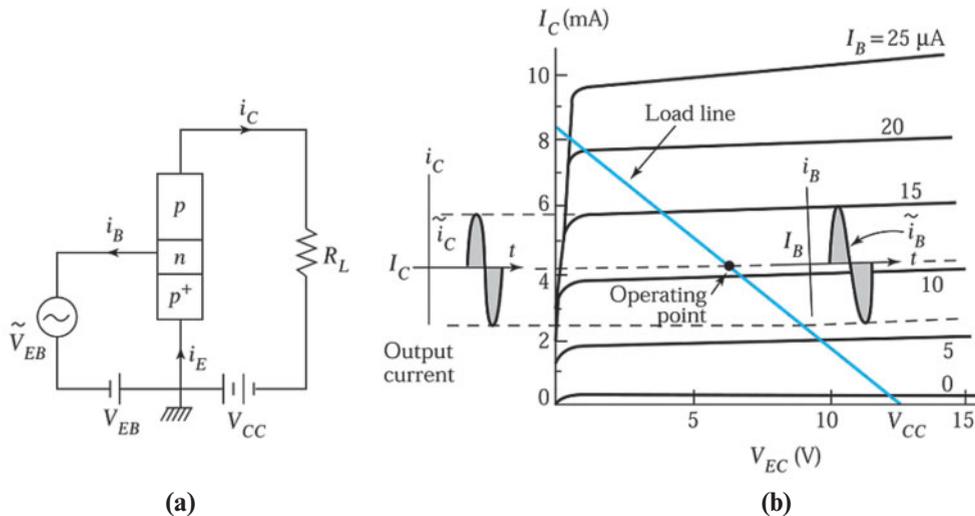


Fig. 12 (a) Bipolar transistor connected in the common-emitter configuration. (b) Small-signal operation of the transistor circuit.

The equivalent circuit for this low-frequency amplifier is shown in Fig. 13a. At higher frequencies, we extend the equivalent circuit by adding the appropriate capacitances. Since the emitter-base junction is forward biased, we expect to have a depletion capacitance C_{EB} and a diffusion capacitance C_d similar to that of a forward-biased $p-n$ junction. For the reverse-biased collector-base junction, we expect to have only a depletion capacitance C_{CB} . The high-frequency equivalent circuit with the three added capacitances is shown in Fig. 13b. Note that $g_m (\equiv \tilde{i}_c / \tilde{v}_{EB})$ is called the *transconductance* and $g_{EB} (\equiv \tilde{i}_B / \tilde{v}_{EB})$ is called the *input conductance*. To take into account the base width modulation effect, there is a finite output conductance $g_{EC} \equiv \tilde{i}_c / \tilde{v}_{EC}$. In addition, we have a base resistance r_B and a collector resistance r_C . Figure 13c represents the high-frequency equivalent circuit incorporating all of the elements.

Cutoff Frequency

In Fig. 13c, the transconductance g_m and the input conductance g_{EB} are dependent on the common-base current gain. At low frequencies, the current gain is a constant, independent of the operating frequency. However, the current gain will decrease after a critical frequency is reached. A typical plot of the current gain versus operating frequency is shown in Fig. 14. The common-base current gain α can be described as

$$\alpha = \frac{\alpha_0}{1 + j(f / f_\alpha)} \quad (38)$$

where α_0 is the low-frequency (or dc) common-base current gain and f_α is the *common-base cutoff frequency*. At $f = f_\alpha$ the magnitude of α is $0.707\alpha_0$ (3 dB down).

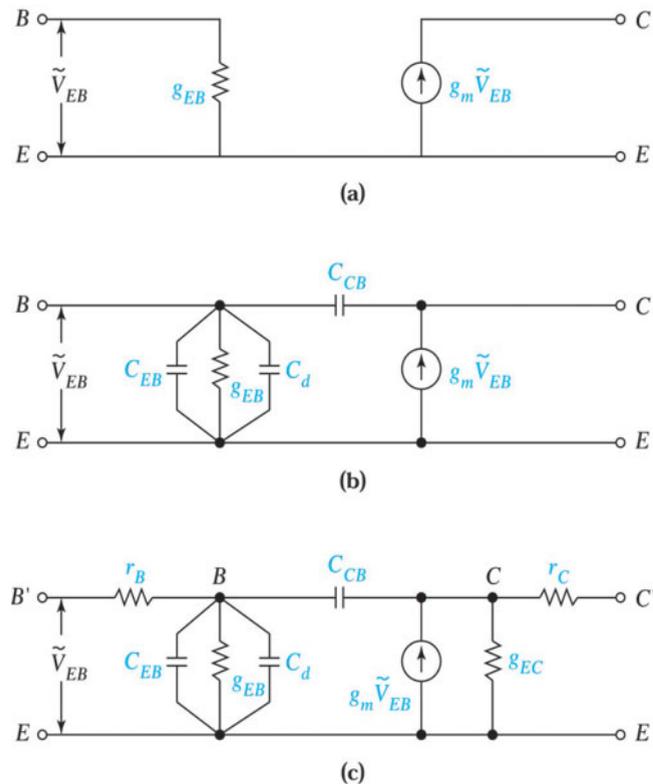


Fig. 13 (a) Basic transistor equivalent circuit. (b) Basic circuit with the addition of depletion and diffusion capacitances. (c) Basic circuit with the addition of resistance and conductance.

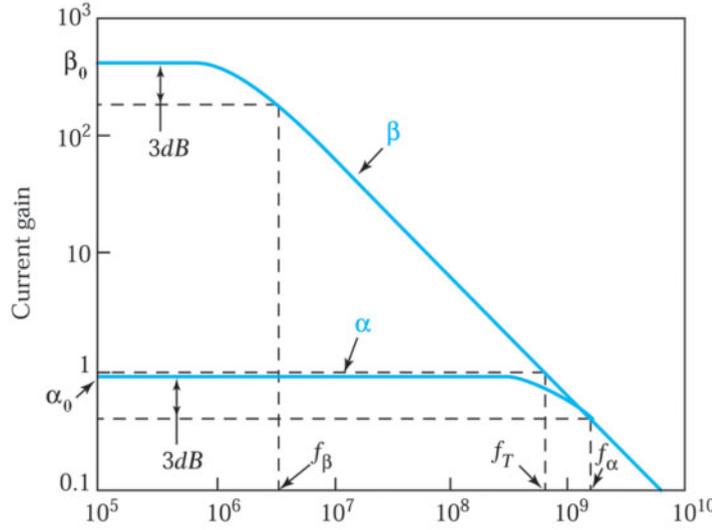


Fig. 14 Current gain as a function of operating frequency.

Figure 14 also shows the common-emitter current gain β . From Eq. 38 we have

$$\beta \equiv \frac{\alpha}{1-\alpha} = \frac{\beta_0}{1+j(f/f_\beta)}, \quad (39)$$

where the f_β is the *common-emitter cut off frequency* and is given by

$$f_\beta = (1-\alpha_0)f_\alpha. \quad (40)$$

Since $\alpha_0 \approx 1$, f_β is much smaller than f_α . Another cutoff frequency is f_T when $|\beta|$ becomes unity. By setting the magnitude of the right-hand side of Eq. 39 equal to 1, we obtain

$$f_T = \sqrt{\beta_0^2 - 1}f_\beta \cong \beta_0(1-\alpha_0)f_\alpha \cong \alpha_0 f_\alpha. \quad (41)$$

Therefore, f_T is very close to but smaller than f_α .

The cutoff frequency f_T can also be expressed as $(2\pi\tau_T)^{-1}$, where τ_T is the total time of the carrier transit from the emitter to the collector. τ_T includes the emitter delay time τ_E , the base transit time τ_B , and the collector transit time τ_C . The most important delay time is τ_B . The distance traveled by the minority carriers in the base in a time interval dt is $dx = v(x)dt$, where $v(x)$ is the effective minority-carrier velocity in the base. This velocity is related to the current as

$$I_p = qv(x)p(x)A, \quad (42)$$

where A is the device area and $p(x)$ is the distribution of the minority carriers. The transit time τ_B required for a hole to traverse the base is given by

$$\tau_B = \int_0^w \frac{dx}{v(x)} = \int_0^w \frac{q p(x)A}{I_p} dx \quad (43)$$

For a straight-line hole distribution, as given by Eq. 15, the integration of Eq. 43 using Eq. 21 for I_p leads to

$$\tau_B = \frac{W^2}{2D_p}. \quad (44)$$

To improve the frequency response, the transit time of minority carriers across the base must be short. Therefore, high-frequency transistors are designed with a narrow base width. Because the electron diffusion constant in silicon is about three times larger than that of holes, all high-frequency silicon transistors are of the n - p - n type (i.e., the minority carrier in the base is the electron). Another way to reduce the base transit time is to use a graded base with a built-in field. For a large doping variation (i.e., high base doping near the emitter and low base doping near the collector), the built-in field in the base helps move carriers toward the collector and reduces the base transit time.

4.3.2 Switching Transients

In digital applications, a transistor is designed to function as a switch. In these applications, we use a small base current to change the collector current from an *off* condition to an *on* condition (or vice versa) in a very short time. The off condition corresponds to a high-voltage and low-current state, and the on condition corresponds to a low-voltage and high-current state. A basic setup of a switching circuit is shown in Fig. 15a, where the emitter-base voltage V_{EB} is suddenly changed from a negative value to a positive value. The output current of the transistor is shown in Fig. 15b. The collector current is initially very low because both the emitter-base junction and the collector-base junction are reverse biased. The current will follow the load line through the active region and will finally reach a high current level, where both junctions become forward biased. Thus, the transistor is virtually open-circuited between the emitter and collector terminals in the *off* condition, which corresponds to the cutoff mode, and short-circuited in the *on* condition, which corresponds to the saturation mode. Therefore, a transistor operated in this mode can nearly duplicate the function of an ideal switch.

The switching time is the time required for a transistor to switch from the off condition to the on condition, or vice versa. Figure 16a shows that when a positive input current pulse is applied to the emitter-base terminal at time $t = 0$, the transistor starts to turn on. At $t = t_2$, the base current is suddenly switched to zero and the transistor starts to turn off. The transient behavior of the collector current I_C can be determined by the variation of the total excess minority carrier charge stored in the base, $Q_B(t)$. A plot of $Q_B(t)$ as a function of time is shown in Fig. 16b. During the turn-on transient, the base-stored charge will increase from zero to $Q_B(t_2)$. During the turn-off transient, the base-stored charge will decrease from $Q_B(t_2)$ to zero. For $Q_B(t) < Q_S$, where Q_S is the base charge when $V_{CB} = 0$ (i.e., at the edge of saturation, as shown in Fig. 16d), the transistor is in the active mode.

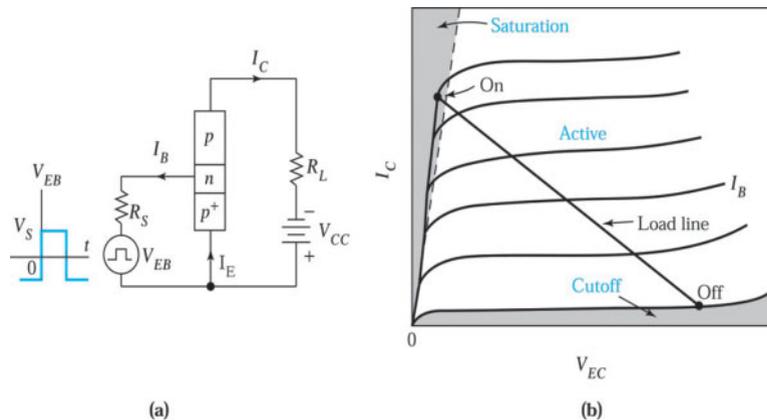


Fig. 15 (a) Schematic of a transistor switching circuit. (b) Switching operation from cutoff to saturation.

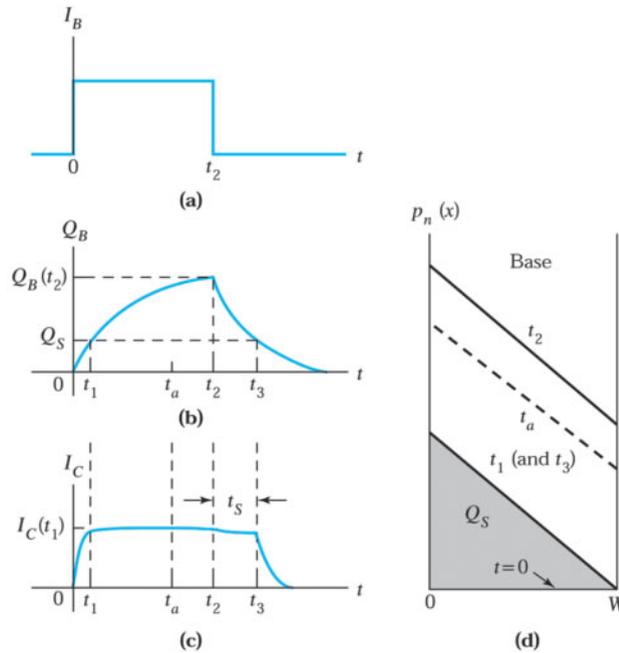


Fig. 16 Transistor switching characteristics. (a) Input base current pulse. (b) Variation of the base-stored charge with time. (c) Variation of the collector current with time. (d) Minority-carrier distributions in the base at different times.

The variation of I_C with time is plotted in Fig. 16c. In the turn-on transient, the stored base charge reaches Q_S , the charge at the edge of saturation at $t = t_1$. For $Q_B > Q_S$ the device is operated in saturation mode, and both the emitter and collector currents remain essentially constant. Figure 16d shows that for any $t > t_1$ (say $t = t_a$), the hole distribution $p_n(x)$ will be parallel to that for $t = t_1$. Therefore, the gradients at $x = 0$ and $x = W$, as well as the currents, remain the same. In the turn-off transient, since the device is initially in the saturation mode, the collector current remains relatively unchanged until Q_B is reduced to Q_S (Fig. 16d). The time from t_2 to t_3 when $Q_B = Q_S$ is called the *storage time delay* t_s . When $Q_B = Q_S$, the device enters the active mode at $t = t_3$. After that time, the collector current will decay exponentially toward zero.

The turn-on time depends on how fast we can add holes (minority carriers in the p - n - p transistor) to the base region. The turn-off time depends on how fast we can remove the holes by recombination. One of the most important parameters for switching transistors is the minority carrier lifetime τ_p . One effective method to reduce τ_p for faster switching is to introduce efficient generation-recombination centers near the midgap.

► 4.4 NONIDEAL EFFECTS

From the above discussion, the emitter doping should be much higher than that of base to have higher current gain. What will it be if the emitter doping is degenerate? In an ideal transistor, the impurity distribution in the base region is uniform. What will it be in a real transistor? The current is limited to the low-level injection described in previous sections. What will it be under a high-injection condition?

4.4.1 Emitter Bandgap Narrowing

To improve current gain, the emitter should be much more heavily doped than the base, that is, $N_E \gg N_B$. However, as the emitter doping becomes very high, we have to consider the bandgap-narrowing effect.³ The bandgap narrowing in heavily doped silicon has been studied based on the broadening of both the conduction band and the valence band in Section 1.6.2 (Chapter 1). Figure 17 shows the experimental data and empirical fit for bandgap narrowing in silicon.⁵ It is approximately analogous to HBT discussed in the following

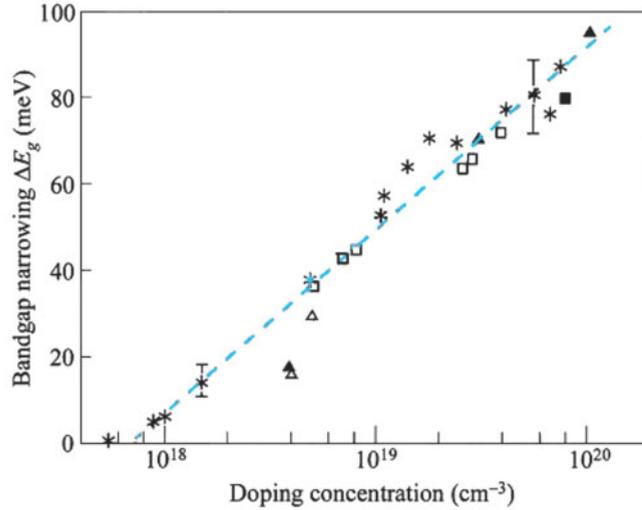


Fig. 17 Experimental data and empirical fit for bandgap narrowing in silicon.⁵

section, and the current gain is reduced with the reduction of emitter bandgap:

$$\beta_0 \sim \frac{N_E}{N_B} \exp\left(-\frac{\Delta E_g}{kT}\right) \quad (45)$$

where ΔE_g is the bandgap reduction of the emitter.

4.4.2 Graded-Base Region

In a real $p-n-p$ transistor fabricated by diffusion or by the ion implantation process of dopant into an epitaxial substrate, the impurity distribution in the base is not uniform but is strongly graded, as shown in Fig. 18a. The corresponding band diagram is shown in Fig. 18b. Because of the impurity gradient, the electrons within the base tend to diffuse toward the collector. However, in thermal equilibrium a built-in electric field exists in the neutral base to counterbalance the diffusion current: that is, the electric field will push the electrons toward the emitter and no current will flow. The same electric field can aid the motion of injected holes. Under an active biasing condition, the injected minority carriers (holes) will move not only by diffusion but also by drift caused by the built-in field of the base region.

The main advantage of the built-in field is to reduce the time needed for the injected holes to travel across the base region. This in turn will improve the transistor's high-frequency response. An associated advantage is the improvement of the base transport factor α_T , since the holes will spend, on the average, less time in the base region and thus will be less likely to recombine with electrons there.

4.4.3 Current Crowding

The base resistance consists of two parts. One is from the base contacts to the emitter edge; the other is the resistance under the emitter shown in Fig. 19,³ which causes a resistive voltage drop and reduces the net V_{BE} across the junction along the emitter edge, and is more severe toward the center of the emitter. The base current I_B decreases with the position toward the center of the emitter. In other words, the emitter current crowds toward the edge of the emitter and increases at high current. The current crowding will cause high injection effects that tend to reduce the current gain.

This current crowding puts some restriction on the design of the emitter strip width. In modern transistors, the emitter strip width can be made small and current crowding is not a major problem. For power transistor, two bases or even an interdigitated structure are used to reduce the base resistance, which, in turn, reduces the current crowding effects.

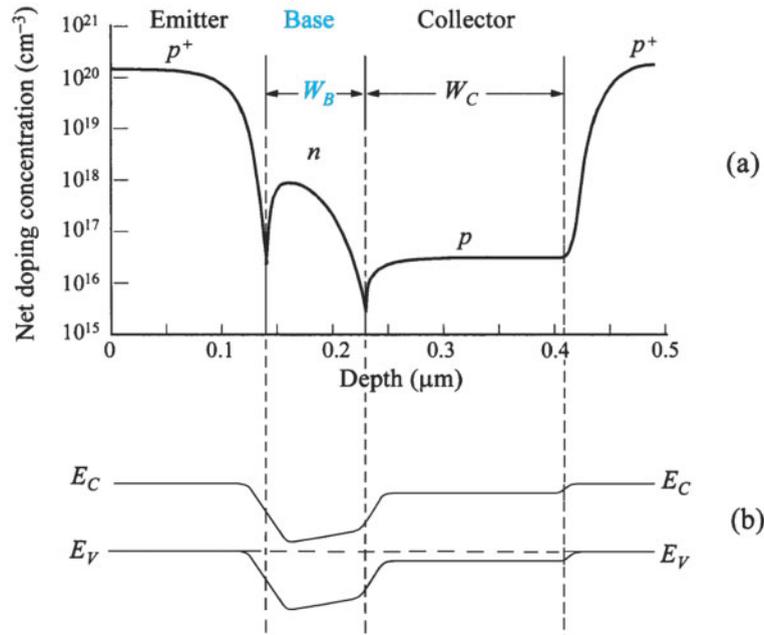


Fig. 18 (a) Impurity distribution in a diffused p - n - p bipolar transistor. (b) Corresponding band diagram in thermal equilibrium.

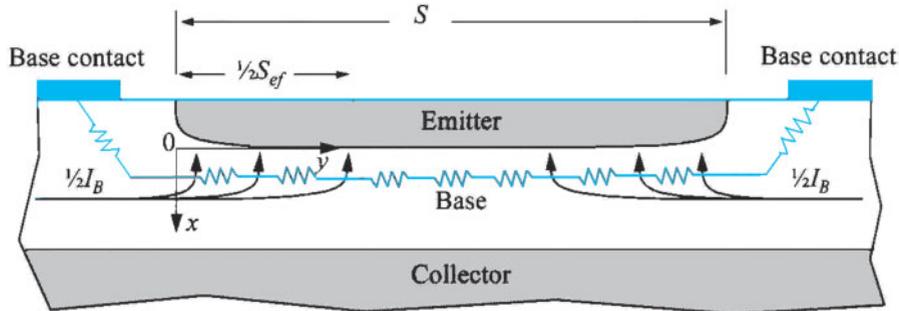
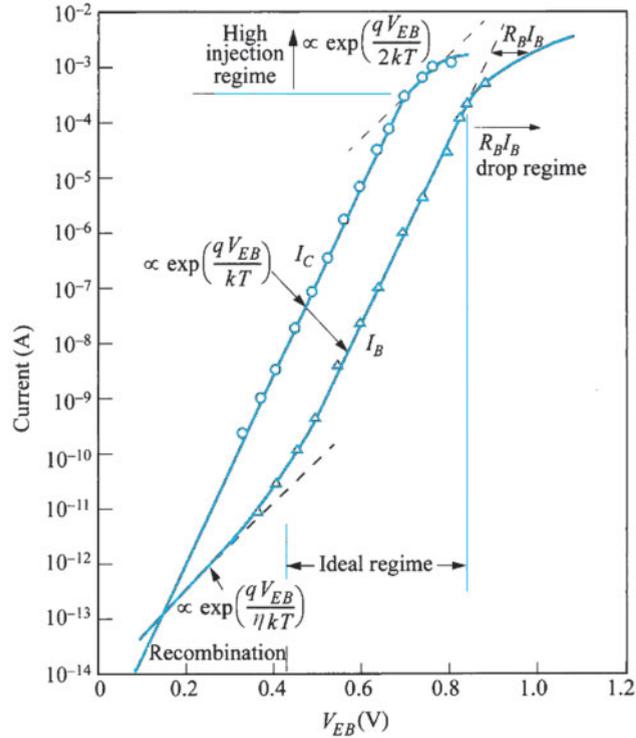


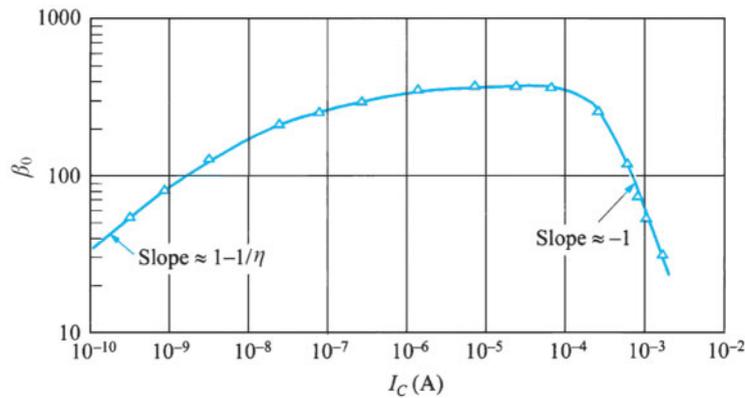
Fig. 19 Cross-section of a two-sided base contact, showing current crowding at high base current. S_{ef} is the effective emitter width.

4.4.4 Generation-Recombination Current and High-Current Effect

In a real transistor, there is a generation current in the depletion region of the reverse-biased base-collector junction. This current is added to the leakage current. A forward-biased emitter-base junction has a recombination current in its depletion region and, therefore, a current component is added to the base current. This recombination current has a profound effect on the current gain. Figure 20a shows the collector current and base current versus V_{EB} for a bipolar transistor operated in the active mode. At low-current levels the recombination current is the dominant current component, and the base current varies as $\exp(qV_{EB}/\eta kT)$, with $\eta \cong 2$. Note that the collector current I_C is not affected by the emitter-base recombination current because I_C is primarily due to those holes injected into the base that diffuse to the collector.



(a)



(b)

Fig. 20 (a) Collector current and base current as functions of emitter-base voltage. (b) Common-emitter current gain for the transistor data in (a).

Figure 20b shows the common-emitter current gain β_0 , which is obtained from Fig. 20a. At low collector current levels, the contribution of the recombination current in the emitter-base depletion region is larger than the diffusion current of minority carriers across the base, so that the emitter efficiency is low. By minimizing the recombination-generation centers in the device, β_0 can be improved at low-current levels. As the base diffusion current becomes dominant, β_0 increases to a high plateau.

At higher collector current levels, β_0 starts to decrease. This is caused by the high-injection effect, where the injected minority carrier density (holes) in the base approaches the impurity concentration and the injected carriers effectively increase the base doping, in turn causing the emitter efficiency to decrease. Another factor contributing to the degradation of β_0 at high-current levels is emitter crowding, which gives rise to a nonuniform distribution of current density under the emitter. The current density at the emitter periphery may be much higher than the average current density. Therefore, the high-injection effect occurs at the emitter periphery, resulting in a reduction of β_0 . In addition, the significant voltage drop on the base resistance also contributes the drop of the current gain at high-injection regime.

► 4.5 HETEROJUNCTION BIPOLAR TRANSISTORS

We have considered the heterojunction in Section 3.7. A heterojunction bipolar transistor (HBT) is a transistor in which one or both p - n junctions are formed between dissimilar semiconductors. The primary advantage of an HBT is its high emitter efficiency (γ). The circuit applications of the HBT are essentially the same as those of bipolar transistors. However, the HBT has higher-speed and higher-frequency capability in circuit operation. Because of these features, the HBT has gained popularity in photonic, microwave, and digital applications. For example, in microwave applications, HBTs are used in solid-state microwave and millimeter-wave power amplifiers, oscillators, and mixers.

4.5.1 Current Gain in HBT

Let semiconductor 1 be the emitter and semiconductor 2 be the base of an HBT. We now consider the impact of the bandgap difference between these two semiconductors on the current gain of an HBT.

When the base-transport factor α_T is very close to unity, the common-emitter current gain can be expressed from Eqs. 8 and 35 as

$$\beta_0 \equiv \frac{\alpha_o}{1 - \alpha_o} \equiv \frac{\gamma \alpha_T}{1 - \gamma \alpha_T} = \frac{\gamma}{1 - \gamma} \quad (\text{for } \alpha_T = 1). \quad (46)$$

Substituting γ from Eq. 31 in Eq. 46 yields (for n - p - n transistors)

$$\beta_0 = \frac{1}{\frac{D_E p_{EO} W}{D_n n_{po} L_E}} \approx \frac{n_{po}}{p_{EO}} \quad (47)$$

The minority carrier concentrations in the emitter and the base are given by

$$p_{EO} = \frac{n_i^2(\text{emitter})}{N_E(\text{emitter})} = \frac{N_C N_V \exp(-E_{gE} / kT)}{N_E}, \quad (48)$$

$$n_{po} = \frac{n_i^2(\text{base})}{N_B(\text{base})} = \frac{N'_C N'_V \exp(-E_{gB} / kT)}{N_B}, \quad (49)$$

where N_C and N_V are the densities of states in the conduction band and the valence band, respectively, and E_{gE} is the bandgap of the emitter semiconductor. N'_C , N'_V , and E_{gB} are the corresponding parameters for the base semiconductor. Therefore, assuming $N_C N_V = N'_C N'_V$

$$\beta_0 \sim \frac{N_E}{N_B} \exp\left(\frac{E_{gE} - E_{gB}}{kT}\right) = \frac{N_E}{N_B} \exp\left(\frac{\Delta E_g}{kT}\right). \quad (50)$$

► EXAMPLE 4

An HBT has a bandgap of 1.62 eV for the emitter and a bandgap of 1.42 eV for the base. A BJT has a bandgap of 1.42 eV for both the emitter and base materials, an emitter doping of 10^{18} cm^{-3} , and a base doping of 10^{15} cm^{-3} .

(a) If the HBT has the same dopings as the BJT, find the improvement of β_0 . (b) If the HBT has the same emitter doping and the same β_0 as the BJT, how much can we increase the base doping of the HBT? Assume that all other device parameters are the same.

SOLUTION

$$(a) \frac{\beta_0(\text{HBT})}{\beta_0(\text{BJT})} = \frac{\exp\left(\frac{E_{gE} - E_{gB}}{kT}\right)}{1} = \exp\left(\frac{1.62 - 1.42}{0.0259}\right) = \exp\left(\frac{0.2}{0.0259}\right) = \exp(7.722) = 2257.$$

We have an improvement of 2257 times in β_0 .

$$(b) \beta_0(\text{HBT}) = \frac{N_E}{N'_B} \exp(7.722) = \beta_0(\text{BJT}) = \frac{N_E}{N_B}$$

$$\therefore N'_B = N_B \exp(7.722) = 2257 \times 10^{15} = 2.26 \times 10^{18} \text{ cm}^{-3}.$$

The base doping of the heterojunction can be increased to $2.26 \times 10^{18} \text{ cm}^{-3}$ to maintain the same β_0 . ◀

4.5.2 Basic HBT Structures

Most developments of HBT technology are for the $\text{Al}_x\text{Ga}_{1-x}\text{As}/\text{GaAs}$ material system. Figure 21a shows a schematic structure of a basic n - p - n HBT. In this device, the n -type emitter is formed in the wide bandgap $\text{Al}_x\text{Ga}_{1-x}\text{As}$, whereas the p -type base is formed in the lower bandgap GaAs. The n -type collector and n -type subcollector are formed in GaAs with light doping and heavy doping, respectively. To facilitate the formation of ohmic contacts, a heavily doped n -type GaAs layer is formed between the emitter contact and the AlGaAs layer. Due to the large bandgap difference between the emitter and the base materials, the common-emitter current gain can be extremely large. However, in homojunction bipolar transistors, there is essentially no bandgap difference; instead, the ratio of the doping concentration in the emitter and base must be very high. This is the fundamental difference between the homojunction and the heterojunction bipolar transistors (see Ex. 4).

Figure 21b shows the energy band diagram of the HBT under the active mode of operation. The bandgap difference between the emitter and the base will provide band offsets at the heterointerface. In fact, the superior performance of the HBT results directly from the valence-band discontinuity ΔE_V at the heterointerface. ΔE_V increases the valence-band barrier height in the emitter-base heterojunction and thus reduces the injection of holes from the base to the emitter. This effect in the HBT allows the use of a heavily doped base while maintaining a high emitter efficiency and current gain. The heavily doped base can reduce the base sheet resistance.⁶

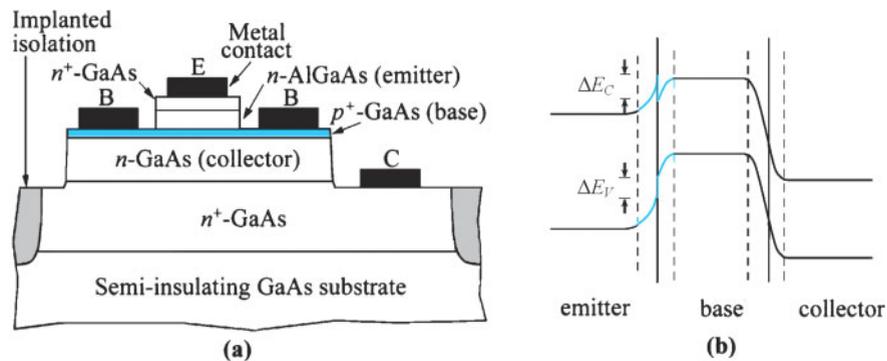


Fig. 21 (a) Schematic cross section of an n - p - n heterojunction bipolar transistor (HBT) structure. (b) Energy band diagram of a HBT operated under active mode.

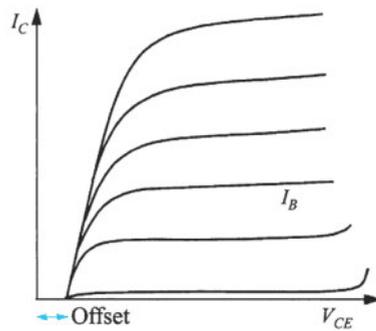


Fig. 22 I_C - V_{CE} characteristics show an offset voltage.

In addition, the base can be made very thin without concern about the *punch-through* effect in the narrow base region. The punch-through effect arises when the base-collector depletion region completely penetrates the base and reaches the emitter-base depletion region. A thin base region is desirable because it reduces the base transit time and increases the cutoff frequency.⁷

One drawback of an HBT is the offset voltage in the common-emitter configuration, as shown in Fig. 22. The offset voltage ΔV_{CE} is defined as the collector-emitter voltage at which the collector current reaches zero. It comes from the potential barrier in the conduction band between the base-emitter junction that hinders the carrier flow into the base and, thus, creates an offset voltage, i.e., an additional voltage drop in the base-emitter junction. The effect of this offset voltage can be alleviated with a graded base-emitter junction, which will be discussed later. It can also be eliminated by incorporating another heterojunction as the base-collector junction.

4.5.3 Advanced HBTs

In recent years the InP-based (InP/InGaAs or AlInAs/InGaAs) material systems have been extensively studied. The InP-based heterostructure has several advantages.⁸ The InP/InGaAs structure has very low surface recombination and, because of a higher electron mobility in InGaAs than in GaAs, superior high-frequency performance is expected. A very high cutoff frequency of 550 GHz is obtained.⁹ In addition, the InP collector region has higher drift velocity at high fields than the GaAs collector. Also, the InP collector breakdown voltage is higher than that of GaAs.

Another heterojunction is in the Si/SiGe material system. This system has several properties that are attractive for HBT applications. Like AlGaAs/GaAs HBTs, Si/SiGe HBTs have high-speed capability since the base can be heavily doped because of the bandgap difference. The small trap density at the silicon surface minimizes the surface recombination current and ensures a high current gain even at low collector current. Compatibility with the standard silicon technology is another attractive feature. Compared with GaAs- and InP-based HBTs, the Si/SiGe HBT, however, has a lower cutoff frequency of 300 GHz¹⁰ because of the lower mobilities in Si.

The conduction band discontinuity ΔE_c shown in Fig. 21b is not desirable, since the discontinuity will make it necessary for the carriers in the heterojunction to transport by means of thermionic emission across a barrier or by tunneling through it. Therefore, the emitter efficiency and the collector current will suffer. The problem can be alleviated by improved structures such as the graded-layer and the graded-base heterojunctions. Figure 23 shows an energy band diagram in which the ΔE_c is eliminated by a graded layer placed between the emitter and base heterojunction. The thickness of the graded layer is W_g .

The base region can also have a graded profile, which results in a reduction of the bandgap from the emitter side to the collector side. The energy band diagram of the graded base HBT is illustrated in Fig. 23 (dotted line). Note that there is a built-in electric field \mathcal{E}_{bi} in the quasi-neutral base. It results in a reduction in the minority-carrier transit time, and thus an increase in the common-emitter current gain and the cutoff frequency of the HBT. \mathcal{E}_{bi} can be obtained, for example, by varying linearly the Al mole fraction x of $\text{Al}_x\text{Ga}_{1-x}\text{As}$ in the base from $x = 0.1$ to $x = 0$.

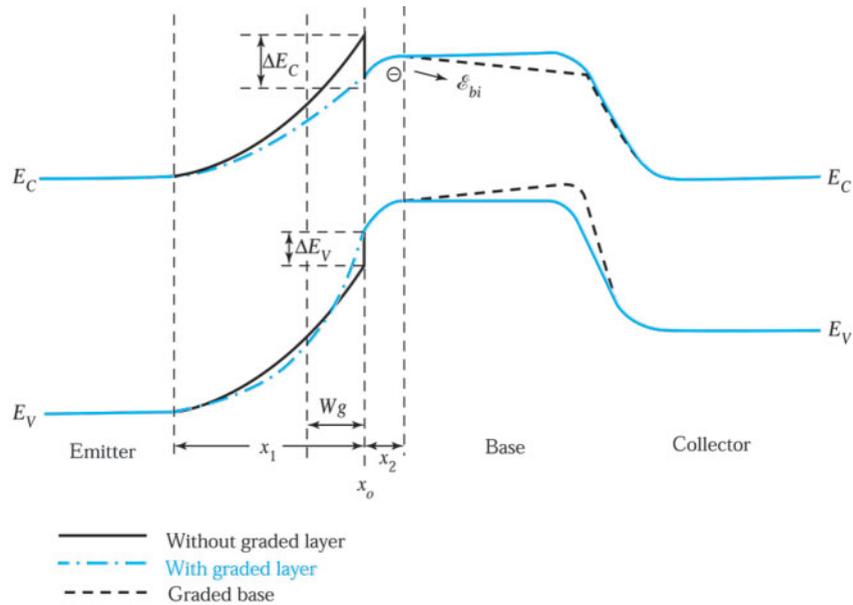


Fig. 23 Energy band diagrams for a heterojunction bipolar transistor with and without a graded layer in the junction and with and without a graded-base layer.

For the design of the collector layer, it is necessary to consider the collector transit time delay and the breakdown voltage requirement. A thicker collector layer will improve the breakdown voltage of the base-collector junction but proportionally increase the transit time. In most devices for high-power applications, the carriers move through the collector at their saturation velocities because very large electric fields are maintained in this layer.

It is possible, however, to increase the velocities by lowering the electric field with certain doping profile in the collector layer. One way is to use p^- collectors with a p^+ pulse-doped layer near the subcollector for an $n-p-n$ HBT. Therefore, electrons entering the collector layer can maintain their higher mobility of the lower valley during most of the collector transit time. Such a device is called a *ballistic collection transistor* (BCT).¹¹ An energy band diagram of a BCT is shown in Fig. 24. The BCT has been shown to have more favorable frequency response characteristics compared with conventional HBTs over a narrow range of bias voltages. Because of its advantages at relatively low collector voltage and current conditions, the BCT is used for switching applications and microwave-power amplifications.

► 4.6 THYRISTORS AND RELATED POWER DEVICES

The thyristor is an important power device that is designed for handling high voltages and large currents. The thyristor is mainly used for switching applications that require the device to change from an *off* or blocking state to an *on* or conducting state, or vice versa.¹² We have considered the use of bipolar transistors in this application, in which the base current drives the transistor from cutoff to saturation for the on-state, and from saturation to cutoff for the off-state. The operation of a thyristor is intimately related to the bipolar transistor, in which both electrons and holes are involved on the transport processes. However, the switching mechanisms in a thyristor are different from those of a bipolar transistor. Also, because of the device construction, thyristors have a much wider range of current- and voltage-handling capabilities. Thyristors are now available¹³ with current ratings from a few milliamperes to over 5000 A and voltage ratings extending above 10,000 V. We first consider the operating principles of basic thyristors and discuss some related high-power and high-frequency thyristors.

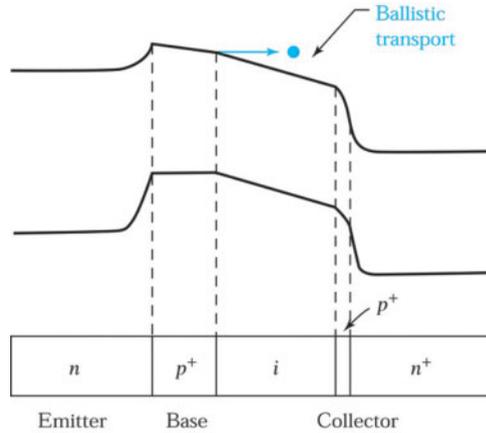


Fig. 24 Energy band diagram for the ballistic collector transistor (BCT).¹¹

4.6.1 Basic Characteristics

Figure 25a shows a schematic cross-sectional view of a thyristor structure, which is a four-layer $p-n-p-n$ device with three $p-n$ junctions in series: J_1 , J_2 , and J_3 . The contact electrode to the outer p -layer is called the *anode* and that to the outer n -layer is called the *cathode*. This structure without any additional electrode is a two-terminal device and is called the $p-n-p-n$ diode. If an additional electrode, called the *gate* electrode, is connected to the inner p -layer (p_2), the resulting three-terminal device is commonly called the *semiconductor-controlled rectifier* (SCR) or *thyristor*.

A typical doping profile of a thyristor is shown in Fig. 25b. An n -type, high-resistivity silicon wafer is chosen as the starting material (n_1 -layer). A diffusion step is used to form the p_1 - and p_2 -layers simultaneously. Finally, an n -type layer is alloyed (or diffused) into one side of the wafer to form the n_2 -layer. Figure 25c shows the energy band diagram of a thyristor in thermal equilibrium. Note that at each junction there is a depletion region with a built-in potential that is determined by the impurity doping profile.

The basic current-voltage characteristics of a $p-n-p-n$ diode are shown in Fig. 26. It exhibits five distinct regions:

- 0-1:** The device is in the forward-blocking or off-state and has very high impedance. Forward breakover (or switching) occurs where $dV/dI = 0$, and at point 1 we define a forward-breakover voltage V_{BF} and a switching current I_S .
- 1-2:** The device is in a negative-resistance region, that is, the current increases as the voltage decreases sharply.
- 2-3:** The device is in the forward-conducting or on state and has low impedance. At point 2, where $dV/dI = 0$, we define the holding current I_h and holding voltage V_h .
- 0-4:** The device is in the reverse-blocking state.
- 4-5:** The device is in the reverse-breakdown region.

Thus, a $p-n-p-n$ diode operated in the forward region is a bistable device that can switch from a high-impedance, low-current off state to a low-impedance, high-current on state and vice versa.

To understand the forward-blocking characteristics, we consider the device as two bipolar transistors, a $p-n-p$ transistor and an $n-p-n$ transistor, connected in a special way, as shown in Fig. 27. They are connected with the base of one transistor attached to the collector of the other, and vice versa. The relationships among the emitter, collector, and base currents and the dc common-base current gain were given in Eqs. 3 and 10. The base current of the $p-n-p$ transistor (transistor 1 with current gain α_1) is

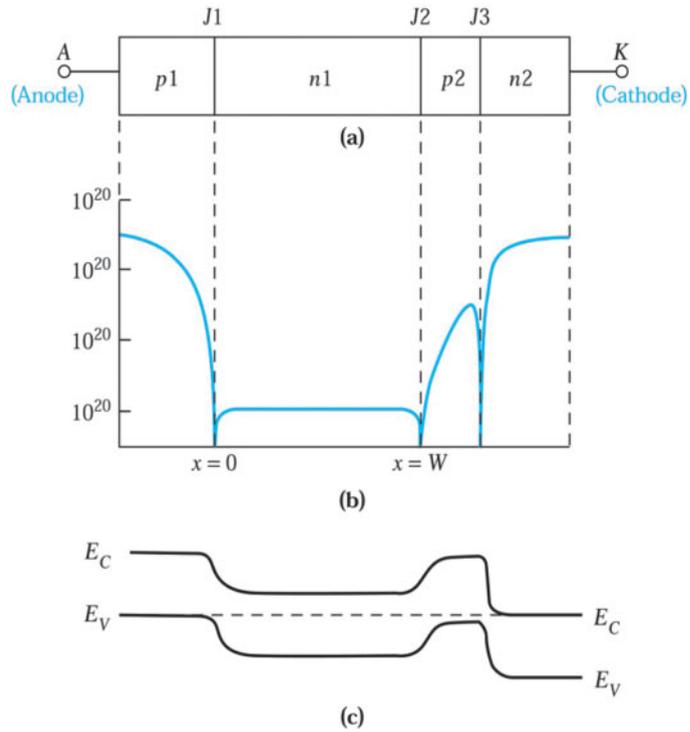


Fig. 25 (a) Four-layer $p-n-p-n$ diode. (b) Typical doping profile of a thyristor. (c) Energy band diagram of a thyristor in thermal equilibrium.

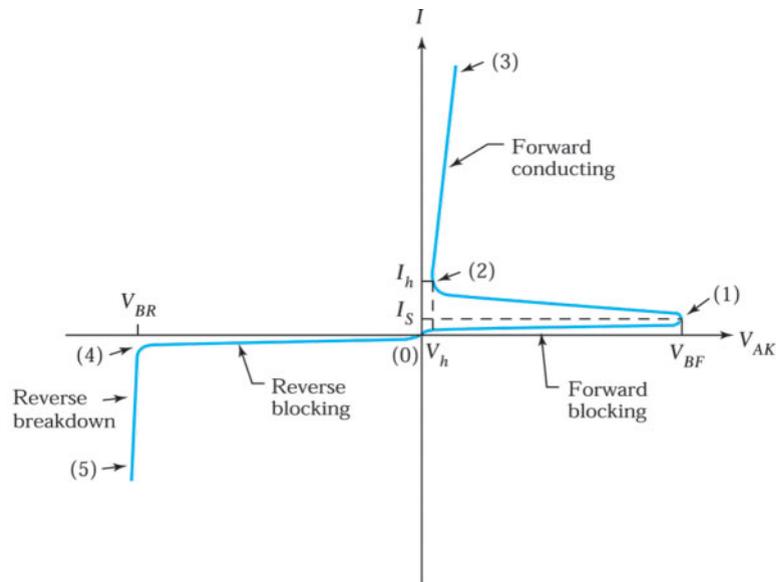


Fig. 26 Current-voltage characteristics of a $p-n-p-n$ diode.

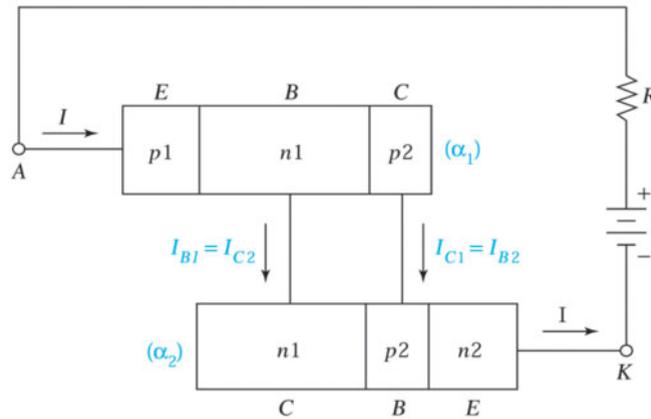


Fig. 27 Two-transistor representation of a thyristor.²

$$\begin{aligned}
 I_{B1} &= I_{E1} - I_{C1} \\
 &= (1 - \alpha_1)I_{E1} - I_1 \\
 &= (1 - \alpha_1)I - I_1,
 \end{aligned} \tag{51}$$

where I_1 is the leakage current I_{CBO} for the transistor 1. This base current is supplied by the collector of the n - p - n transistor (transistor 2 with the current gain α_2). The collector current of the n - p - n transistor is

$$I_{C2} = \alpha_2 I_{E2} + I_2 = \alpha_2 I + I_2, \tag{52}$$

where I_2 is the leakage current I_{CBO} for the transistor 2. By equating I_{B1} and I_{C2} , we obtain

$$I = \frac{I_1 + I_2}{1 - (\alpha_1 + \alpha_2)}. \tag{53}$$

► EXAMPLE 5

Consider a thyristor in which the leakage currents I_1 and I_2 are 0.4 and 0.6 mA, respectively. Explain the forward-blocking characteristics when $(\alpha_1 + \alpha_2)$ is 0.01 and 0.9999.

SOLUTION The current gains are functions of the current I and generally increase with increasing current. At low currents both α_1 and α_2 are much less than 1, and we have

$$I = \frac{0.4 \times 10^{-3} + 0.6 \times 10^{-3}}{1 - 0.01} = 1.01 \text{ mA}$$

In this case, the current flowing through the device is the sum of the leakage currents I_1 and I_2 ($\cong 1$ mA). As the applied voltage increases, the current I also increases, as do α_1 and α_2 . This in turn causes I to increase further—a regenerative behavior. When $\alpha_1 + \alpha_2 = 0.9999$,

$$I = \frac{0.4 \times 10^{-3} + 0.6 \times 10^{-3}}{1 - 0.9999} = 10 \text{ A.}$$

This value is 10,000 times larger than $I_1 + I_2$. Therefore, as $(\alpha_1 + \alpha_2)$ approaches 1, the current I increases without limit; that is, the device is at forward conduction. ◀

The variations of the depletion layer widths of a $p-n-p-n$ diode biased in different regions are shown in Fig. 28. At thermal equilibrium, Fig. 28a, there is no current flowing and the depletion layer widths are determined by the impurity doping profiles. In the forward-blocking state, Fig. 28b, junction $J1$ and $J3$ are forward biased and $J2$ is reverse biased. Most of the voltage drop occurs across the central junction $J2$. In the forward-conduction state, Fig. 28c, all three junctions are forward biased. The two transistors ($p1-n1-p2$ and $n1-p2-n2$) are in saturation mode of operation. Therefore, the voltage drop across the device is very low, given by $(V_1 - |V_2| + V_3)$, which is approximately equal to the voltage drop across one forward-biased $p-n$ junction. In the reverse-blocking state, Fig. 28d, junction $J2$ is forward biased but both $J1$ and $J3$ are reversed biased. For the doping profile shown in Fig. 25b, the reverse-breakdown voltage will be mainly determined by $J1$ because of the lower impurity concentration in the $n1$ -region.

Figure 29a shows the device configuration of a thyristor that is fabricated by planar processes with a gate electrode connected to the $p2$ -region. A cross section of the thyristor along the dashed lines is shown in Fig. 29b. The current-voltage characteristic of the thyristor is similar to that of the $p-n-p-n$ diode, except that the gate current I_g causes an increase of $\alpha_1 + \alpha_2$ and results in a breakover at a lower voltage. Figure 30 shows the effect of gate current on the current-voltage characteristics of a thyristor. As the gate current increases, the forward breakover voltage decreases.

A simple application of a thyristor is shown in Fig. 31a, where a variable power is delivered to a load from a constant line source. The load R_L may be a light bulb or a heater, such as furnace. The amount of power delivered to the load during each cycle depends on the timing of the gate-current pulses of the thyristor (Fig. 31b). If the current pulses are delivered to the gate near the beginning of each cycle, more power will be delivered to the load. However, if the current pulses are delayed, the thyristor will not turn on until later in the cycle, and the amount of power delivered to the load will be substantially reduced.

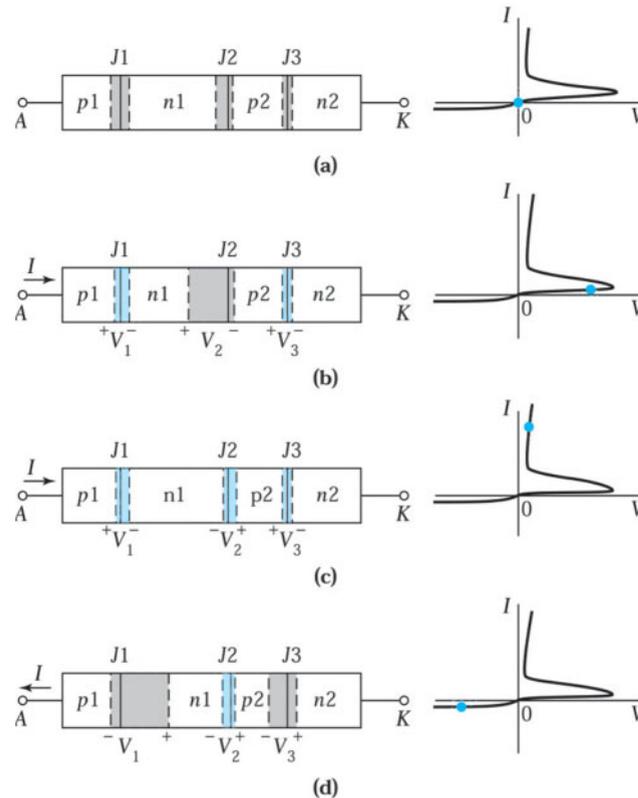


Fig. 28 Depletion layer widths and voltage drops of a thyristor operated under (a) equilibrium, (b) forward blocking, (c) forward conducting, and (d) reverse blocking.

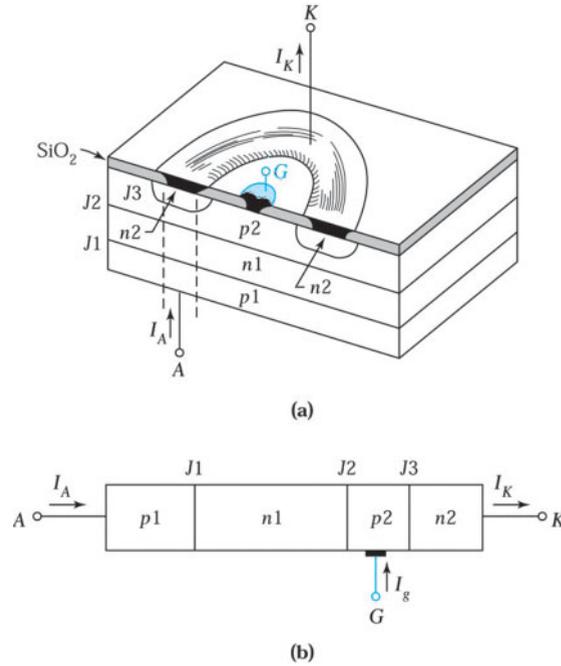


Fig. 29 (a) Schematic of a planar three-terminal thyristor. (b) One-dimensional cross section of the planar thyristor.

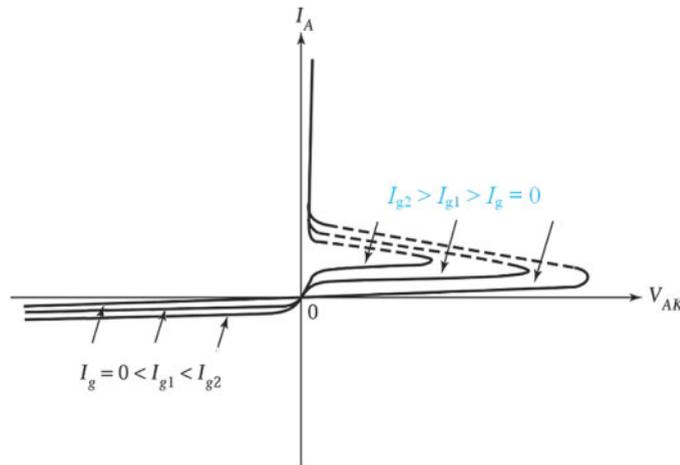


Fig. 30 Effect of gate current on current-voltage characteristics of a thyristor.

4.6.2 Bidirectional Thyristor

A bidirectional thyristor is a switching device that has on and off states for positive and negative anode voltages and is therefore useful in ac applications. The bidirectional $p-n-p-n$ diode switch is called a *diac* (diode ac switch). It behaves like two conventional $p-n-p-n$ diodes with the anode of the first diode connected to the cathode of the second, and vice versa. Figure 32a illustrates such a structure where $M1$ stands for main terminal 1 and $M2$ for main terminal 2. When we integrate this arrangement into a single two-terminal device, we have a diac, as shown in Fig. 32b. The symmetry of this structure will result in identical performance for either polarity of applied voltage.

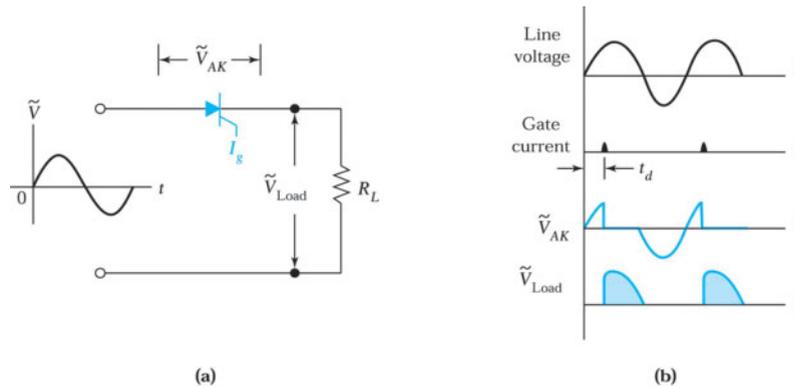


Fig. 31 (a) Schematic circuit for a thyristor application. (b) Wave forms of voltages and gate current.

When a positive voltage is applied to $M1$ with respect to $M2$, junction $J4$ is reverse biased so that the $n2'$ region does not contribute to the functioning of the device. Therefore, the $p1-n1-p2-n2$ layers constitute a $p-n-p-n$ diode that produces the forward portion of the $I-V$ characteristic shown in Fig. 32c. If a positive voltage is applied to $M2$ with respect to $M1$, a current will conduct in the opposite direction and $J3$ will be reverse biased. Therefore, the $p1'-n1'-p2'-n2'$ layers of the reverse $p-n-p-n$ diode produce the reverse portion of the $I-V$ characteristics shown in Fig. 32c.

A bidirectional three-terminal thyristor is called a *triac* (*triode ac* switch). The triac can switch the current in either direction by applying a low-voltage, low-current pulse of either polarity between the gate and one of the two main terminals, $M1$ and $M2$, as shown in Fig. 33. The operational principles and the $I-V$ characteristics of a triac are similar to those of a diac. By adjusting the gate current, the breakover voltage can be varied in either polarity.

► SUMMARY

The bipolar transistor, developed in 1947, remains one of the most important semiconductor devices. A bipolar transistor is formed when two $p-n$ junctions of the same semiconductor materials are physically close enough to interact. Charge carriers injected from the forward-biased first junction result in a large current flow in the reverse-biased second junction.

We have considered the static characteristics of bipolar transistors, such as the modes of operation and the current-voltage characteristics of the common-emitter configuration. We have also considered the frequency response and switching behavior. A key device parameter of a bipolar transistor is the base width, which must be very small compared with the minority-carrier diffusion length to improve the current gain and to increase the cutoff frequency.

Bipolar transistors are used extensively as discrete devices or in integrated circuits for current-gain, voltage-gain, and power-gain applications. They are also used in bipolar-CMOS combination circuits (BiCMOS), covered in Chapters 6 and 15, for high-density, high-speed operations.

The frequency limitations of a conventional bipolar transistor are the result of its low-base doping and relatively wide base. To overcome these limitations, a heterojunction bipolar transistor (HBT) formed between two dissimilar semiconductors can have much higher-base doping and a much narrower base. The HBT has, therefore, gained popularity in millimeter-wave and high-speed digital applications.

Another important bipolar device is the thyristor, which is formed of three or more $p-n$ junctions. The thyristor is used mainly for switching applications. These devices can have current ratings from a few milliamperes to over 5000 A and voltage ratings extending above 10,000 V. We considered the basic characteristics of thyristor operation. In addition, we discussed the bidirectional thyristor (diac and triac) that has on-off states with either positive or negative terminal voltages. Thyristors can cover a wide range of applications from low-frequency high-current power supplies to high-frequency, low-power applications, including lighting controls, home appliances, and industrial equipment.

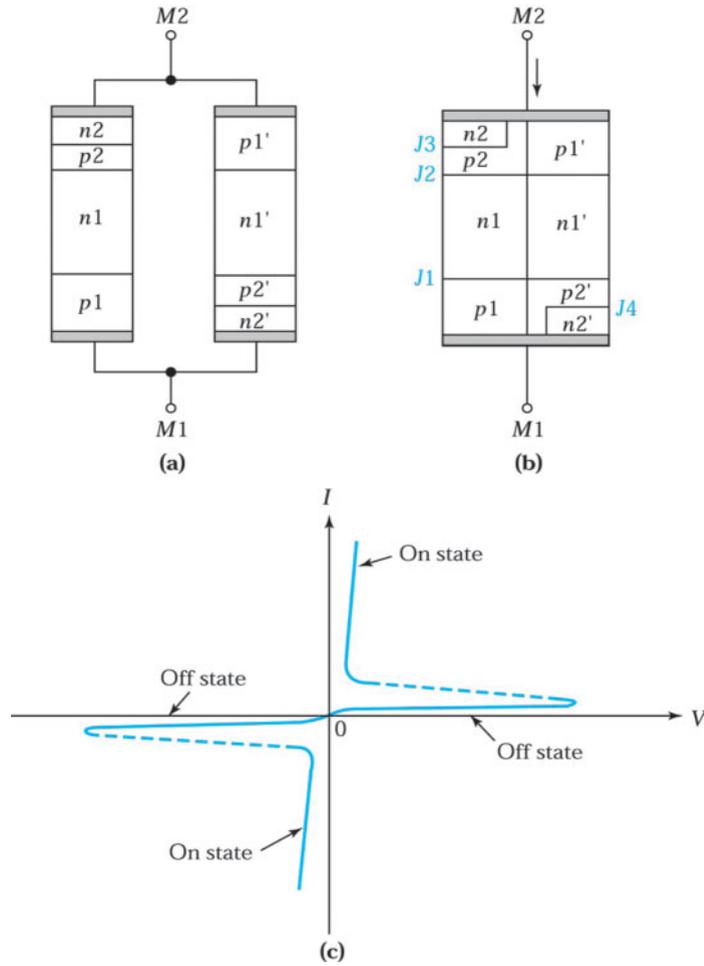


Fig. 32 (a) Two reverse-connected p - n - p - n diodes. (b) Integration of the diodes into a single two-terminal diode ac switch (diac). (c) Current-voltage characteristics of a diac.

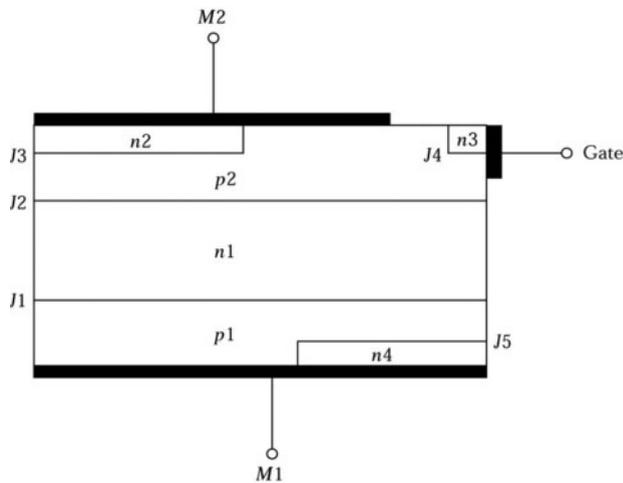


Fig. 33 Cross section of a triode ac switch (triac), a six-layer structure having five p - n junctions.

► REFERENCES

1. (a) J. Bardeen and W. H. Brattain, “The Transistor, A Semiconductor Triode,” *Phys. Rev.*, **74**, 230 (1948). (b) W. Shockley, “The Theory of $p-n$ Junction in Semiconductors and $p-n$ Junction Transistor,” *Bell Syst. Tech. J.*, **28**, 435 (1949).
2. J. J. Ebers, “Four-Terminal $p-n-p-n$ Transistor,” *Proc. IEEE*, **40**, 1361 (1952).
3. S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, 3rd Ed., Wiley Interscience, Hoboken, 2007.
4. J. M. Early, “Effects of Space-Charge Layer Widening in Junction Transistors,” *Proc. IRE*, **40**, 1401 (1952).
5. J. del Alamo, S. Swirhum, and R. M. Swanson, “Simultaneous Measurement of Hole Lifetime, Hole Mobility and Bandgap Narrowing in Heavily Doped n -Type Silicon,” *Tech. Dig. IEEE IEDM*, 290 (1985).
6. J. S. Yuan and J. J. Liou, “Circuit Modeling for Transient Emitter Crowding and Two-Dimensional Current and Charge Distribution Effects,” *Solid-State Electron.*, **32**, 623 (1989).
7. J. J. Liou, “Modeling the Cutoff Frequency of Heterojunction Bipolar Transistors Subjected to High Collector-Layer Currents,” *J. Appl. Phys.*, **67**, 7125 (1990).
8. B. Jalali and S. J. Pearton, Eds., *InP HBTs: Growth, Processing, and Application*, Artech House, Norwood, 1995.
9. W. Hafez and M. Feng, “0.25 μm Emitter InP SHBTs with $f_T = 550$ GHz and $BV_{CEO} > 2\text{V}$,” *Tech. Dig. IEEE Int. Electron Devices Meet.*, p. 549 (2004).
10. N. Zerounian et al., “500 GHz Cutoff Frequency SiGe HBTs,” *Electron. Lett.*, **43**, 774 (2007).
11. T. Ishibashi and Y. Yamauchi, “A Possible Near-Ballistic Collection in an AlGaAs/GaAs HBT with a Modified Collector Structure,” *IEEE Trans. Electron Devices*, **ED-35**, 401 (1988).
12. P. D. Taylor, *Thyristor Design and Realization*, Wiley, New York, 1993.
13. H. P. Lips, “Technology Trends for HVDC Thyristor Valves,” *1998 Int. Conf. Power Syst. Tech. Proc.*, **1**, 446 (1998).

► PROBLEMS (* INDICATES DIFFICULT PROBLEMS)

FOR SECTION 4.2 STATIC CHARACTERISTICS OF BIPOLAR TRANSISTORS

1. An $n-p-n$ transistor has a base transport factor α_T of 0.998, an emitter efficiency of 0.997, and an I_{C_0} of 10 nA. (a) Calculate α_0 and β_0 for the device. (b) If $I_B = 0$, what is the emitter current?
2. Given that an ideal transistor has an emitter efficiency of 0.999 and the collector-base leakage current is 10 μA , calculate the active region emitter current due to holes if $I_B = 0$.
3. A silicon $p-n-p$ transistor has impurity concentrations of 5×10^{18} , 2×10^{17} , and 10^{16}cm^{-3} in the emitter, base, and collector, respectively. The base width is 1.0 μm and the device cross-sectional area is 0.2 mm^2 . When the emitter-base junction is forward biased to 0.5 V and the base-collector junction is reverse biased to 5 V, calculate (a) the neutral base width and (b) the minority carrier concentration at the emitter-base junction.

4. For the transistor in Prob. 3, the diffusion constants of minority carriers in the emitter, base, and collector are 52, 40, and 115 cm²/s, respectively, and the corresponding lifetimes are 10⁻⁸, 10⁻⁷, and 10⁻⁶ s. Find the current components I_{Ep} , I_{Cp} , I_{En} , I_{Cn} , and I_{BB} illustrated in Fig. 5.
5. Using the results obtained from Prob. 3 and 4, (a) find the terminal currents I_E , I_C , and I_B of the transistor; (b) calculate emitter efficiency, base transport factor, common-base current gain, and common-emitter current gain; and (c) comment on how the emitter efficiency and base transport factor can be improved.
6. Referring to the minority carrier concentration shown in Eq. 14, sketch $p_n(x)/p_n(0)$ curves as a function of x with different W/L_p . Show that the distribution will approach a straight line when W/L_p is small enough (say $W/L_p < 0.1$).
- *7. For a transistor under the active mode of operation, use Eq. 14 to find the exact solutions of I_{Ep} and I_{Cp} .
8. Derive the expression for total excess minority-carrier charge Q_B , if the transistor is operated under the active mode and $p_n(0) \gg p_{no}$. Explain how the charge can be approximated by the triangle area in the base shown in Fig. 6. In addition, using the parameters in Prob. 3, find Q_B .
9. Using Q_B derived from Prob. 8, show that the collector current expressed in Eq. 27 can be approximated by $I_C \cong (2D_p/W^2)Q_B$.
10. Show that the base transport factor α_T can be simplified to $1 - (W^2/2L_p^2)$.
11. If the emitter efficiency is very close to unity, show that the common-emitter current gain β_0 can be given by $2L_p^2/W^2$. (Hint: Use α_T in Prob. 10.)
12. For a p^+n-p transistor with high emitter efficiency, find the common-emitter current gain β_0 . If the base width is 2 μm and the diffusion constant of minority carrier in the base region is 100 cm²/s, assume that the lifetime of the carrier in the base region is 3×10^{-7} s. (Hint: Refer to β_0 derived in Prob. 11.)
13. A silicon $n-p-n$ bipolar transistor has impurity concentrations of 3×10^{18} , 2×10^{16} , and 5×10^{15} cm⁻³ in the emitter, base, and collector, respectively. Determine the diffusion constants of minority carrier in the three regions by using Einstein's relationship, $D = (kT/q)\mu$. Assume that the mobilities of electrons and holes, μ_n and μ_p , can be expressed as

$$\mu_n = 88 + \frac{1252}{(1 + 0.698 \times 10^{-17} N)} \quad \text{and} \quad \mu_p = 54.3 + \frac{407}{(1 + 0.374 \times 10^{-17} N)} \quad \text{at } T = 300 \text{ K.}$$
14. Using the results obtained from Prob. 13, determine the current components in each region with $V_{BE} = 0.6\text{V}$ (operated under active mode). The device cross-sectional area is 001 mm² and the neutral-base width is 0.5 μm . Assume the minority-carrier lifetime in each region is the same and equal to 10⁻⁶ s.
15. Based on the results obtained from Prob. 14, find the emitter efficiency, base transport factor, common-base current gain, and common-emitter current gain.
16. For an ion implanted $n-p-n$ transistor, the net impurity doping in the neutral base is $N(x) = N_{AO}e^{-x/l}$, where $N_{AO} = 2 \times 10^{18}$ cm⁻³ and $l = 0.3 \mu\text{m}$. (a) Find the total number of impurities in the neutral-base region per unit area. (b) Find the average impurity concentration in the neutral-base region for a neutral-base width of 0.8 μm .
17. Referring to Problem 16, if $L_E = 1 \mu\text{m}$, $N_E = 10^{19}$ cm⁻³, $D_E = 1$ cm²/s, the average lifetime is 10⁻⁶ s in the base, and the average diffusion coefficient in the base corresponds to the impurity concentration in Prob. 16, find the common-emitter current gain.
18. Estimate the collector current level for the transistor in Probs. 16 and 17 that has an emitter area of 10⁻⁴ cm². The base resistance of the transistor can be expressed as $10^{-3} \bar{\rho}_B / W$ where W is the neutral-base width and $\bar{\rho}_B$ is the average base resistivity.

- *19. Plot the common-emitter current gain as a function of the base current I_B from 0 to 25 μA at a fixed V_{EC} of 5 V for the transistor shown in Fig. 10b. Explain why the current gain is not a constant.
20. The general equations of the emitter and collector currents for the basic Ebers-Moll model [J. J. Ebers and J. L. Moll, "Large-Single Behavior of Junction Transistors," *Proc. IRE.*, **42**, 1761 (1954)] are

$$\begin{aligned} I_E &= I_{FO}(e^{qV_{EB}/kT} - 1) - \alpha_R I_{RO}(e^{qV_{CB}/kT} - 1), \\ I_C &= \alpha_F I_{FO}(e^{qV_{EB}/kT} - 1) - I_{RO}(e^{qV_{CB}/kT} - 1), \end{aligned}$$

where α_F and α_R are the *forward common-base current gain* and the *reverse common-base current gain*, respectively. I_{FO} and I_{RO} are the saturation currents of the normally forward- and reverse-biased diodes, respectively. Find α_F and α_R in terms of the constants in Eqs. 25, 26, 28, and 29.

- *21. Referring to the transistor in Example 2, find I_E and I_C by using the equations derived in Problem 20.
22. Derive Eq. 32b for the collector current starting with the field-free steady-state continuity equation. (Hint: Consider the minority carrier distribution in the collector region.)

FOR SECTION 4.3 FREQUENCY RESPONSE AND SWITCHING OF BIPOLAR TRANSISTORS

23. A Si transistor has D_p of 10 cm^2/s and W of 0.5 μm . Find the cutoff frequencies for the transistor with a common-base current gain β_0 of 0.998. Neglect the emitter and collector delays.
24. If we want to design a bipolar transistor with 5 GHz cutoff frequency f_T , what the neutral base width W will be? Assume D_p is 10 cm^2/s and neglect the emitter and collector delays.

FOR SECTION 4.5 HETEROJUNCTION BIPOLAR TRANSISTORS

25. Consider a $\text{Si}_{1-x}\text{Ge}_x/\text{Si}$ HBT with $x = 10\%$ in the base region (and 0% in emitter and collector region). The bandgap of the base region is 9.8% smaller than that of Si. If the base current is due to emitter injection efficiency only, what is the expected change in the common-emitter current gain between 0° and 100°C?
26. For an $\text{AlGa}_{1-x}\text{As}/\text{GaAs}$ HBT, the bandgap of the $\text{Al}_x\text{Ga}_{1-x}\text{As}$ is a function of x and can be expressed as 1.424 + 1.247 x eV (when $x \leq 0.45$) and 1.9 + 0.125 x + 0.143 x^2 eV (when $0.45 < x \leq 1$). Plot $\beta_0(\text{HBT})/\beta_0(\text{BJT})$ as a function of x .

FOR SECTION 4.6 THYRISTORS AND RELATED POWER DEVICES

27. For the doping profile shown in Fig. 25, find the width W ($> 10\mu\text{m}$) of the $n1$ -region so that the thyristor has a reverse blocking voltage of 120 V. If the current gain α_2 for the $n1$ - $p2$ - $n2$ transistor is 0.4 independent of current, and α_1 of the $p1$ - $n1$ - $p2$ transistor can be expressed as $0.5\sqrt{L_p/W} \ln(J/J_0)$, where L_p is 25 μm and J_0 is $5 \times 10^{-6} \text{ A}/\text{cm}^2$, find the cross-sectional area of the thyristor that will switch at a current I_S of 1 mA.

MOS Capacitor and MOSFET

- ▶ 5.1 IDEAL MOS CAPACITOR
- ▶ 5.2 SiO_2 -SI MOS CAPACITOR
- ▶ 5.3 CARRIER TRANSPORT IN MOS CAPACITORS
- ▶ 5.4 CHARGE-COUPLED DEVICES
- ▶ 5.5 MOSFET FUNDAMENTALS
- ▶ SUMMARY

The metal-oxide-semiconductor (MOS) capacitor is of paramount importance in semiconductor device physics because the device is extensively used in the study of semiconductor surfaces.* In integrated circuits, the device is used as storage capacitor and forms the basic building block for charge-coupled devices (CCD). The metal-oxide semiconductor field-effect transistor (MOSFET) is composed of an MOS capacitor and two p - n junctions placed immediately adjacent to the MOS capacitor.¹ Since its first demonstration in 1960, the MOSFET has been developed quickly and has become the most important device for advanced integrated circuits such as microprocessors and semiconductor memories. This is because MOSFET has many unique and unprecedented features, including low-power consumption and a high manufacturing yield.

Specifically, we cover the following topics:

- The ideal and practical MOS capacitors.
- The inversion condition and threshold voltage of an MOS capacitor.
- C-V and I-V characteristics of an MOS capacitor.
- The charge-coupled device.
- Basic characteristics of MOSFET.

▶ 5.1 IDEAL MOS CAPACITOR

A perspective view of an MOS capacitor is shown in Fig. 1*a*. The cross section of the device is shown in Fig. 1*b*, where d is the thickness of the oxide and V is the applied voltage on the metal plate. Throughout this section we use the convention that the voltage V is positive when the metal plate is positively biased with respect to the ohmic contact and V is negative when the metal plate is negatively biased with respect to the ohmic contact.

The energy band diagram of an ideal p -type semiconductor MOS at $V = 0$ is shown in Fig. 2.¹ The work function is the energy difference between the Fermi level and the vacuum level (i.e., $q\phi_m$ for the metal and $q\phi_s$ for the semiconductor). Also shown are the electron affinity $q\chi$, which is the energy difference between the conduction band edge and the vacuum level in the semiconductor, $q\chi_s$, the oxide electron affinity, $q\phi_B$, the energy barrier between the metal and the oxide, and $q\psi_B$, the energy difference between the Fermi level E_F and the intrinsic Fermi level E_i .

*A more general class of device is the metal-insulator-semiconductor (MIS) capacitor. However, because in most experimental studies the insulator has been silicon dioxide, in this text the term “MOS capacitor” is used interchangeably with “MIS capacitor.”

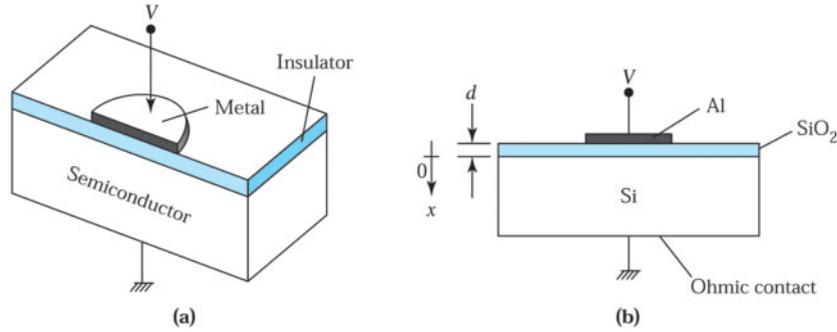


Fig. 1 (a) Perspective view of a metal-oxide-semiconductor (MOS) capacitor. (b) Cross-section of an MOS capacitor.

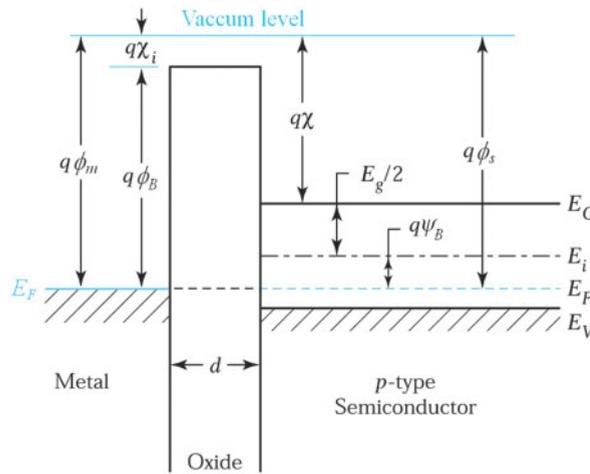


Fig. 2 Energy band diagram of an ideal MOS capacitor at $V = 0$.

An ideal MOS is defined as follows: (a) At zero applied bias, the energy difference between the metal work function $q\phi_m$ and the semiconductor work function $q\phi_s$ is zero, or the work function difference $q\phi_{ms}$ is zero: *

$$q\phi_{ms} \equiv (q\phi_m - q\phi_s) - q\phi_m \left(q\chi + \frac{E_g}{2} + q\psi_B \right) = 0, \quad (1)$$

where the sum of the three items in the brackets equals $q\phi_s$. In other words, the energy band is flat (flat-band condition) when there is no applied voltage. (b) The only charges that exist in the capacitor under any biasing conditions are those in the semiconductor and those with equal but opposite sign on the metal surface adjacent to the oxide. (c) There is no carrier transport through the oxide under direct current (dc)-biasing conditions, or the resistivity of the oxide is infinite. This ideal MOS theory serves as a foundation for understanding practical MOS devices.

When an ideal MOS capacitor is biased with positive or negative voltages, three cases can exist at the semiconductor surface. For a p -type semiconductor, when a negative voltage ($V < 0$) is applied to the metal plate, excess positive carriers (holes) will be induced at the SiO_2 -Si interface. In this case, the bands near the semiconductor surface are bent upward, as shown in Fig. 3a. For an ideal MOS capacitor, no current flows in the device regardless of the value of the applied voltage, therefore, the Fermi level in the semiconductor remains

* This is for a p -type semiconductor. For an n -type semiconductor, the term $q\psi_B$ is replaced by $-q\psi_B$.

constant. Previously, we determined that the carrier density in the semiconductor depends exponentially on the energy difference $E_i - E_F$, that is,

$$p_p = n_i e^{(E_i - E_F)/kT} \quad (2)$$

The upward bending of the energy band at the semiconductor surface causes an increase in the energy $E_i - E_F$ there, which in turn gives rise to an enhanced concentration or accumulation of holes near the oxide-semiconductor interface. This is called the *accumulation case*. The corresponding charge distribution is shown on the right side of Fig. 3a, where Q_s is the positive charge per unit area in the semiconductor and Q_m is the negative charge per unit area ($|Q_m| = Q_s$) in the metal. When a small positive voltage ($V > 0$) is applied to an ideal MOS capacitor, the energy bands near the semiconductor surface are bent downward and the majority carriers (holes) are depleted (Fig. 3b). This is called the *depletion case*. The space charge per unit area, Q_{sc} , in the semiconductor is equal to $-qN_A W$, where W is the width of the surface depletion region.

When a larger positive voltage is applied, the energy bands bend downward even more so that the intrinsic level E_i at the surface crosses over the Fermi level, as shown in Fig 3c. As a result, the positive

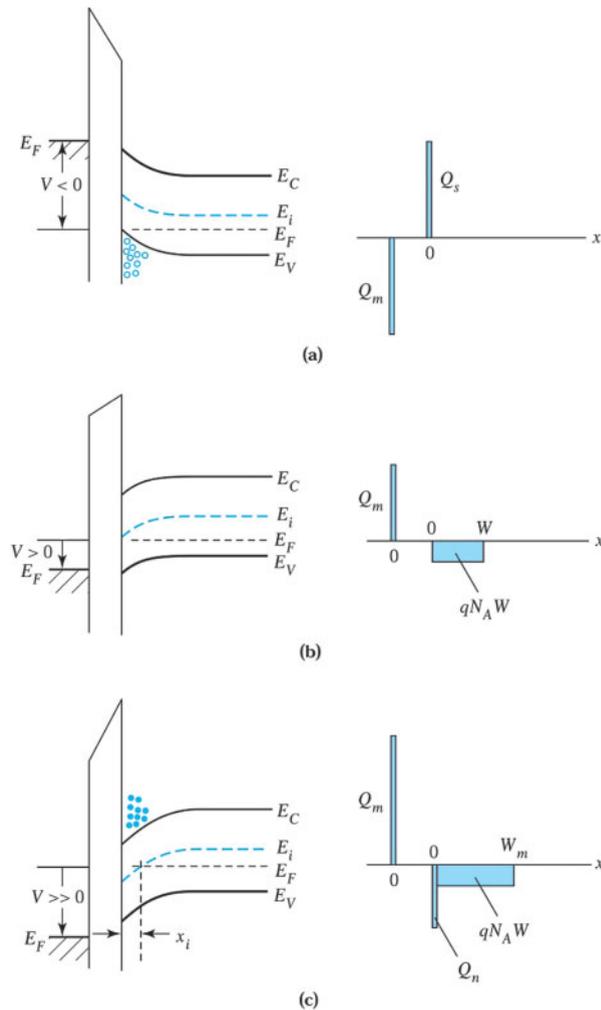


Fig. 3 Energy band diagrams and charge distributions of an ideal MOS capacitor in (a) accumulation, (b) depletion, and (c) inversion cases.

gate voltage starts to induce excess negative carriers (electrons) at the SiO₂-Si interface. The electron concentration in the semiconductor depends exponentially on the energy difference $E_F - E_i$, and is given by

$$n_p = n_i e^{(E_F - E_i)/kT} \quad (3)$$

In the case shown in Fig. 3c, $(E_F - E_i) > 0$. Therefore, the electron concentration n_p at the interface is larger than n_i , and the hole concentration given by Eq. 2 is less than n_i . The number of electrons (minority carriers) at the surface is greater than holes (majority carriers); the surface is thus inverted. This is called the *inversion* case.

Initially, the surface is in a *weak inversion* condition since the electron concentration is small. As the bands are bent further, eventually the conduction band edge comes close to the Fermi level. The onset of *strong inversion* occurs when the electron concentration near the SiO₂-Si interface is equal to the substrate doping level. After this point most of the additional negative charges in the semiconductor consist of the charge Q_n (Fig. 3c) in a very narrow *n*-type inversion layer $0 \leq x \leq x_i$, where x_i is the width of the inversion region. Typically, the value of x_i , ranges from 1 to 10 nm and is always much smaller than the surface depletion-layer width.

Once strong inversion occurs, the surface depletion-layer width reaches a maximum. This is because when the bands are bent downward far enough for strong inversion to occur, even a very small increase in band bending corresponding to a very small increase in depletion-layer width results in a large increase in the charge Q_n in the inversion layer. Thus, under a strong inversion condition the charge per unit area Q_s in the semiconductor is the sum of the charge Q_n in the inversion layer and the charge Q_{sc} in the depletion region:

$$Q_s = Q_n + Q_{sc} = Q_n - qN_A W_m, \quad (4)$$

where W_m is the maximum width of the surface depletion region.

The Surface Depletion Region

Figure 4 shows a more detailed band diagram at the surface of a *p*-type semiconductor. The electrostatic potential ψ is defined as zero in the bulk of the semiconductor. At the semiconductor surface, $\psi = \psi_s$; ψ_s is called the surface potential. We can express electron and hole concentrations in Eqs. 2 and 3 as a function of ψ :

$$n_p = n_i e^{q(\psi - \psi_B)/kT}, \quad (5a)$$

$$p_p = n_i e^{q(\psi_B - \psi)/kT}, \quad (5b)$$

where ψ is positive when the band is bent downward, as shown in Fig. 4. At the surface the densities are

$$n_s = n_i e^{q(\psi_s - \psi_B)/kT}, \quad (6a)$$

$$p_s = n_i e^{q(\psi_B - \psi_s)/kT}. \quad (6b)$$

From this discussion and with the help of Eq. 6, the following regions of surface potential can be distinguished:

| | |
|-----------------------|--|
| $\psi_s < 0$ | Accumulation of holes (bands bend upward). |
| $\psi_s = 0$ | Flat-band condition. |
| $\psi_B > \psi_s > 0$ | Depletion of holes (bands bend downward). |
| $\psi_s = \psi_B$ | Midgap with $n_s = n_p = n_i$ (intrinsic concentration). |
| $\psi_s > \psi_B$ | Inversion (bands bend downward). |

The potential ψ as a function of distance can be obtained by using the one-dimensional Poisson's equation:

$$\frac{d^2\psi}{dx^2} = \frac{-\rho_s(x)}{\epsilon_s}, \quad (7)$$

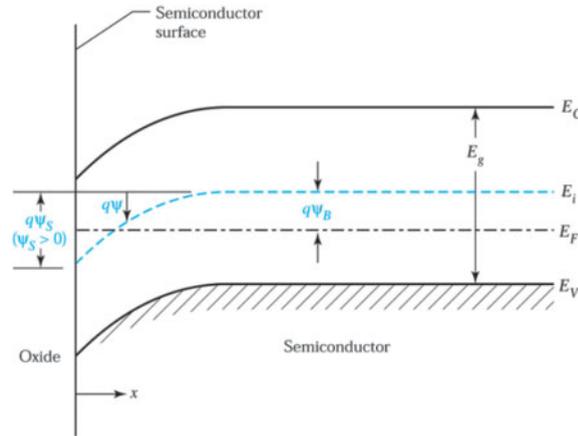


Fig. 4 Energy band diagrams at the surface of a *p*-type semiconductor.

where $\rho_s(x)$ is the charge density per unit volume at position x and ϵ_s is the dielectric permittivity. We use the depletion approximation that we have employed in the study of *p-n* junctions. When the semiconductor is depleted to a width of W and the charge within the semiconductor is given by $\rho_s = -qN_A$, integration of Poisson's equation gives the electrostatic potential distribution as a function of distance x in the surface depletion region:

$$\psi = \psi_s \left(1 - \frac{x}{W} \right)^2 \tag{8}$$

The surface potential ψ_s is

$$\psi_s = \frac{qN_A W^2}{2\epsilon_s} \tag{9}$$

Note that the potential distribution is identical to that for a one-sided *n⁺-p* junction.

The surface is inverted whenever ψ_s is larger than ψ_B . However, we need a criterion for the onset of strong inversion, after which the charges in the inversion layer become significant. A simple criterion is that the electron concentration at the surface is equal to the substrate impurity concentration, i.e., $n_s = N_A$. Since $N_A = n_i e^{q\psi_B/kT}$, from Eq. 6a we obtain

$$\boxed{\psi_s (inv) \cong 2\psi_B = \frac{2kT}{q} \ln \left(\frac{N_A}{n_i} \right)} \tag{10}$$

Equation 10 states that a potential ψ_B is required to bend the energy bands down to the intrinsic condition at the surface ($E_i = E_F$), and bands must then be bent downward by another $q\psi_B$ at the surface to obtain the condition of strong inversion.

As discussed previously, the surface depletion layer reaches a maximum when the surface is strongly inverted. Accordingly, the maximum width of the surface depletion region W_m is given by Eq. 9 in which ψ_s equals $\psi_s(inv)$, or

$$W_m = \sqrt{\frac{2\epsilon_s \psi_s(inv)}{qN_A}} \cong \sqrt{\frac{2\epsilon_s (2\psi_B)}{qN_A}} \tag{11a}$$

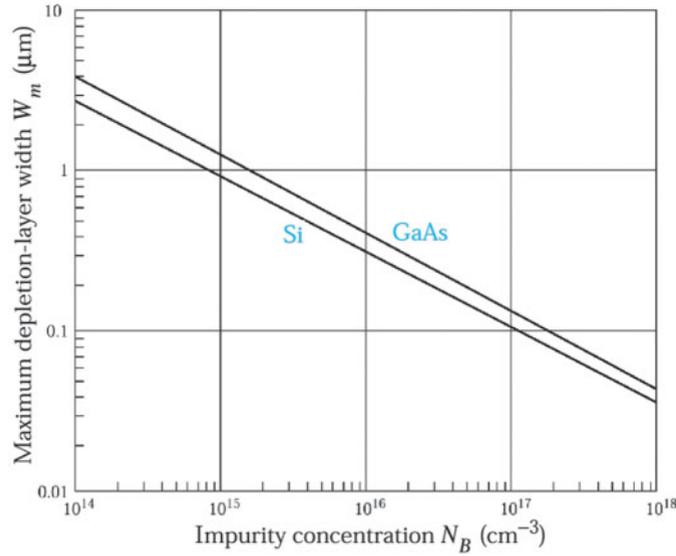


Fig. 5 Maximum depletion-layer width versus impurity concentration of Si and GaAs under strong-inversion condition.

or

$$W_m = 2\sqrt{\frac{\epsilon_s kT \ln\left(\frac{N_A}{n_i}\right)}{q^2 N_A}} \quad (11b)$$

and

$$Q_{sc} = qN_A W_m \cong \sqrt{2q\epsilon_s N_A (2\psi_B)} \quad (12)$$

► EXAMPLE 1

For an ideal metal-SiO₂-Si capacitor having $N_A = 10^{17} \text{ cm}^{-3}$, calculate the maximum width of the surface depletion region.

SOLUTION At room temperature, $kT/q = 0.026 \text{ V}$ and $n_i = 9.65 \times 10^9 \text{ cm}^{-3}$, the dielectric permittivity of Si is $11.9 \times 8.85 \times 10^{-14} \text{ F/cm}$. From Eq. 11b,

$$\begin{aligned} W_m &= 2\sqrt{\frac{11.9 \times 8.85 \times 10^{-14} \times 0.026 \ln\left(\frac{10^{17}}{9.65 \times 10^9}\right)}{1.6 \times 10^{-19} \times 10^{17}}} \\ &= 10^{-5} \text{ cm} = 0.1 \text{ } \mu\text{m} . \end{aligned}$$

The relationship between W_m and the impurity concentration is shown in Fig. 5 for silicon and gallium arsenide, where N_B is equal to N_A for p -type and N_D for n -type semiconductors.

Ideal MOS Curves

Figure 6a shows the energy band diagram of an ideal MOS capacitor with band bending identical to that shown in Fig. 4. The charge distribution is shown in Fig. 6b. In the absence of any work function

differences, the applied voltage will appear partly across the oxide and partly across the semiconductor. Thus,

$$V = V_o + \psi_s, \quad (13)$$

where V_o is the potential across the oxide and is given (Fig. 6c) by

$$V_o = \mathcal{E}_o d = \frac{|Q_s| d}{\epsilon_{ox}} \equiv \frac{|Q_s|}{C_o}, \quad (14)$$

where \mathcal{E}_o is the field in the oxide, Q_s is the charge per unit area in the semiconductor, and $C_o (= \epsilon_{ox}/d)$ is the oxide capacitance per unit area. The corresponding electrostatic potential distribution is shown in Fig. 6d.

The total capacitance C of the MOS capacitor is a series combination (Fig. 7a, inset) of the oxide capacitance C_o and the semiconductor depletion-layer capacitance C_j :

$$C = \frac{C_o C_j}{(C_o + C_j)} \text{ F/cm}^2, \quad (15)$$

where $C_j = \epsilon_s/W$, the same as for an abrupt p - n junction.

From Eqs. 9, 13, 14, and 15, we can eliminate W and obtain the formula for the capacitance:

$$\frac{C}{C_o} = \frac{1}{\sqrt{1 + \frac{2\epsilon_{ox}^2 V}{qN_A \epsilon_s d^2}}}, \quad (16)$$

which predicts that the capacitance will decrease with increasing metal-plate voltage while the surface is being depleted. When the applied voltage is negative, there is no depletion region, and we have an accumulation of holes at the semiconductor surface. As a result, the total capacitance is close to the oxide capacitance ϵ_{ox}/d .

In the other extreme, when strong inversion occurs, the width of the depletion region will not increase with a further increase in applied voltage. This condition takes place at a metal-plate voltage that causes the surface potential ψ_s to reach $\psi_s(inv)$, as given in Eq. 10. Substituting $\psi_s(inv)$ into Eq. 13 and noting that the corresponding charge per unit area is $qN_A W_m$ yields the metal-plate voltage at the onset of strong inversion. This voltage is called the threshold voltage:

$$V_T = \frac{qN_A W_m}{C_o} + \psi_s(inv) \cong \frac{\sqrt{2\epsilon_s qN_A (2\psi_B)}}{C_o} + 2\psi_B. \quad (17)$$

Once the strong inversion takes place, the total capacitance will remain at a minimum value given by Eq. 15 with $C_j = \epsilon_s/W_m$,

$$C_{\min} = \frac{\epsilon_{ox}}{d + \left(\frac{\epsilon_{ox}}{\epsilon_s}\right)W_m}. \quad (18)$$

A typical capacitance-voltage curve of an ideal MOS capacitor is shown in Fig. 7a based on both the depletion approximation (Eqs. 16–18) and exact calculations (solid curve). Note the close correlation between the depletion approximation and the exact calculations.

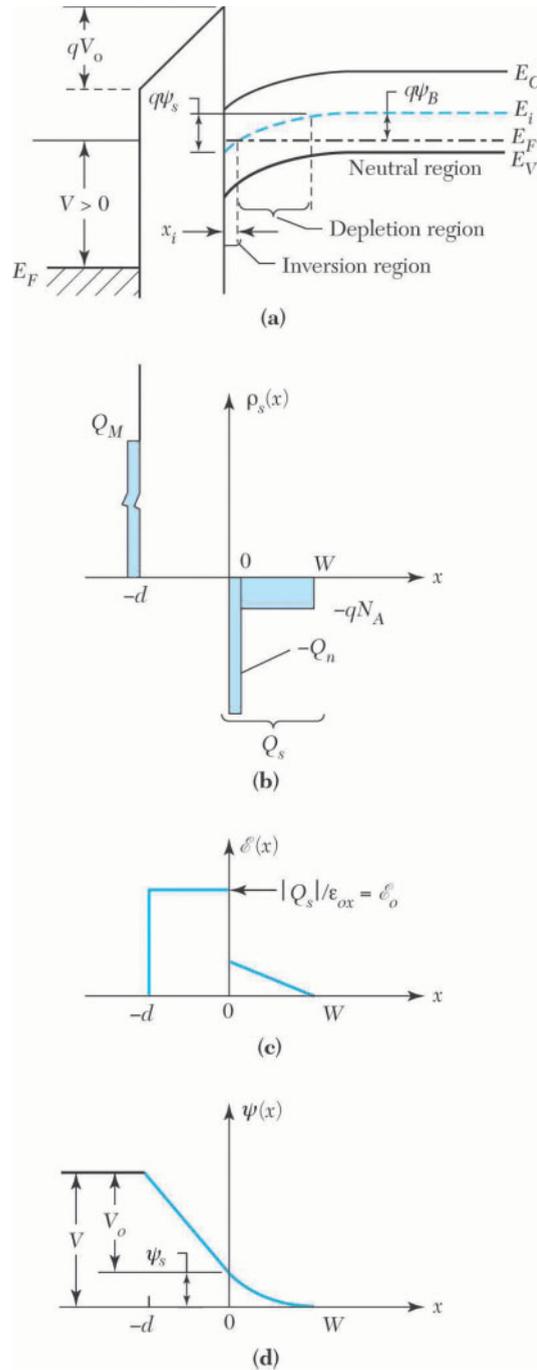


Fig. 6 (a) Band diagram of an ideal MOS capacitor. (b) Charge distributions under inversion condition. (c) Electric-field distribution. (d) Potential distribution.

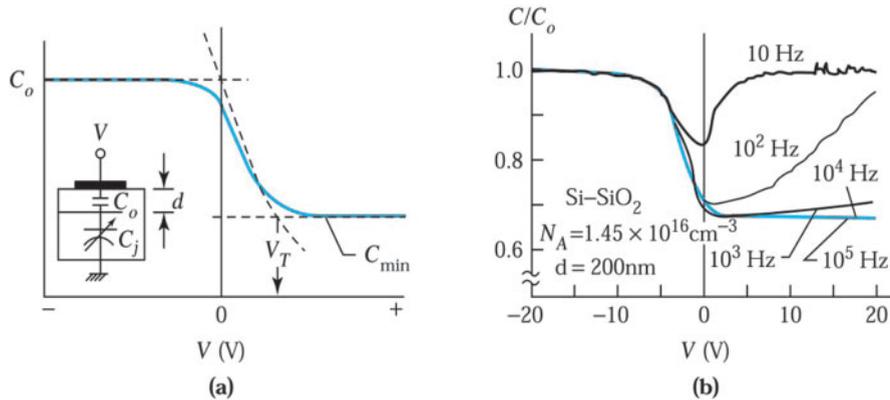


Fig. 7 (a) High-frequency MOS C - V curve showing its approximated segments (dashed lines). Inset shows the series connection of the capacitors. (b) Effect of frequency on the C - V curve.²

Although we have considered only the p -type substrate, all of the considerations are equally valid for an n -type substrate with the proper changes in signs and symbols (e.g., Q_p for Q_n). The capacitance-voltage characteristics will have identical shapes but will be mirror images of each other, and the threshold voltage is a negative quantity for an ideal MOS capacitor on an n -type substrate.

In Fig. 7a we assumed that when the voltage on the metal plate changes, all the incremental charge appears at the edge of the depletion region. Indeed, this happens when the measurement frequency is high. If, however, the measurement frequency is low enough so that generation-recombination rates in the surface depletion region are equal to or greater than the voltage variation, then the electron concentration (minority carrier) can follow the alternating current (ac) signal and lead to charge exchange with the inversion layer in step with the measurement signal. As a result the capacitance in strong inversion will be that of the oxide layer alone, C_o . Figure 7b shows the measured MOS C - V curves at different frequencies.² Note that the onset of the low-frequency curves occurs at $f \leq 100$ Hz.

► EXAMPLE 2

For an ideal metal-SiO₂-Si capacitor having $N_A = 10^{17}$ cm⁻³ and $d = 5$ nm, calculate the minimum capacitance of the C - V curve in Fig. 7a. The dielectric constant of SiO₂ is 3.9.

SOLUTION

$$C_o = \frac{\epsilon_{ox}}{d} = \frac{3.9 \times 8.85 \times 10^{-14}}{5 \times 10^{-7}} = 6.90 \times 10^{-7} \text{ F/cm}^2.$$

$$Q_{sc} = qN_A W_m = 1.6 \times 10^{19} \times 10^{17} \times (1 \times 10^{-5}) = 1.6 \times 10^7 \text{ C/cm}^2.$$

W_m is obtained in Example 1.

$$\psi_s(\text{inv}) \approx 2\psi_B = \frac{2kT}{q} \ln\left(\frac{N_A}{n_i}\right) = 2 \times 0.026 \times \ln\left(\frac{10^{17}}{9.65 \times 10^9}\right) = 0.84 \text{ V}.$$

The minimum capacitance C_{\min} at V_T is

$$C_{\min} = \frac{\epsilon_{ox}}{d + (\epsilon_{ox}/\epsilon_s)W_m} = \frac{3.9 \times 8.85 \times 10^{-14}}{(5 \times 10^{-7}) + (3.9/11.9)(1 \times 10^{-5})} = 9.1 \times 10^{-8} \text{ F/cm}^2.$$

Therefore, C_{\min} is about 13% of C_o .

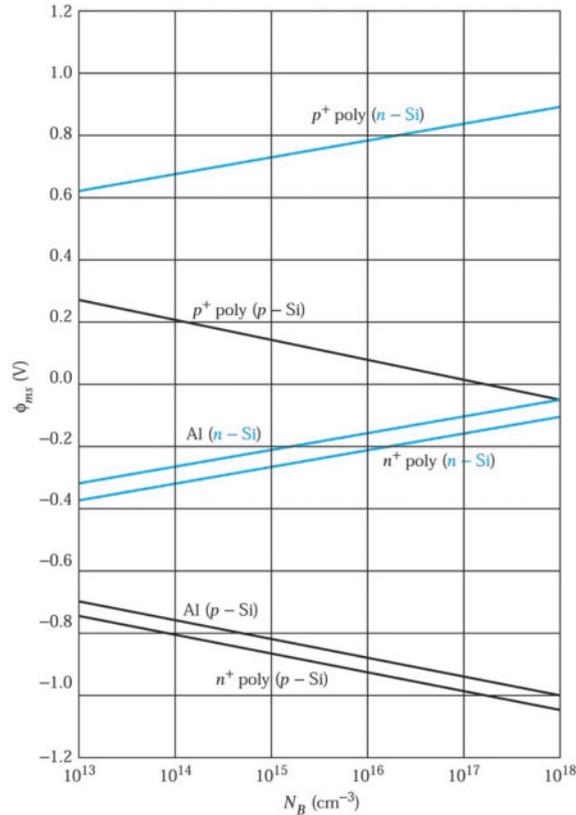


Fig. 8 Work function difference as a function of background impurity concentration for Al, n^+ -, and p^+ - polysilicon gate materials.

► 5.2 SiO₂-SI MOS CAPACITOR

Of all the MOS capacitors, the metal-SiO₂-Si is the most extensively studied. The electrical characteristics of the SiO₂-Si system approach those of the ideal MOS capacitor. However, for commonly used metal electrodes, the work function difference $q\phi_{ms}$ is generally not zero, and there are various charges inside the oxide or at the SiO₂-Si interface that will, in one way or another, affect the ideal MOS characteristics.

The Work Function Difference

The work function of a semiconductor $q\phi_s$, which is the energy difference between the vacuum level and the Fermi level (Fig. 2), varies with the doping concentration. For a given metal with a fixed work function $q\phi_m$ we expect that the work function difference $q\phi_{ms} \equiv (q\phi_m - q\phi_s)$ will vary depending on the doping of the semiconductor. One of the most common metal electrodes is aluminum, with $q\phi_m = 4.1$ eV. Another material also used extensively is the heavily doped polycrystalline silicon (also called polysilicon). The work function for n^+ - and p^+ -polysilicon are 4.05 and 5.05 eV, respectively. Figure 8 shows the work function differences for aluminum, n^+ -, and p^+ -polysilicon on silicon as the doping is varied. It is interesting to note that ϕ_{ms} can vary over a 2 V range depending on the electrode materials and the silicon doping concentration.

To construct the energy band diagram of an MOS capacitor, we start with an isolated metal and an isolated semiconductor with an oxide layer sandwiched between them (Fig. 9a). In this isolated situation, all bands are flat; this is the flat-band condition. At thermal equilibrium, the Fermi level must be a constant and the vacuum level must be continuous. To accommodate the work function difference, the semiconductor bands bend downward, as shown in Fig. 9b. Thus, the metal is positively charged and the semiconductor surface is negatively charged at

thermal equilibrium. To achieve the ideal flat-band condition of Fig. 2, we have to apply a voltage equal to the work function difference $q\phi_{ms}$. This corresponds exactly to the situation shown in Fig. 9a, where we must apply a negative voltage V_{FB} , called the flat-band voltage ($V_{FB} = \phi_{ms}$), to the metal.

Interface Traps and Oxide Charges

In addition to the work function difference, the MOS capacitor is affected by charges in the oxide and traps at the SiO_2 -Si interface. The basic classification of these traps and charges are shown in Fig. 10. They are the interface-trapped charge, fixed-oxide charge, oxide-trapped charge, and mobile ionic charge.³

Interface-trapped charges Q_{it} are the charges in interface traps (historically also called interface states, fast states, or surface states) due to the interruption of the periodic lattice structure at the SiO_2 -Si interface, and are dependent on the chemical composition of this interface. The traps are located at the SiO_2 -Si interface with energy states in the silicon forbidden bandgap. The interface trap density, i.e., number of interface traps per unit area and per eV, is orientation dependent. In $\langle 100 \rangle$ orientation, the interface trap density is about an order of magnitude smaller than that in $\langle 111 \rangle$ orientation. Present-day MOS capacitors with thermally grown silicon dioxide on silicon have most of the interface-trapped charges passivated by low-temperature (450°C) hydrogen annealing. The value of Q_{it}/q for $\langle 100 \rangle$ -oriented silicon can be as low as 10^{10} cm^{-2} , which amounts to about one interface-trapped charge per 10^5 surface atoms. For $\langle 111 \rangle$ -oriented silicon, Q_{it}/q is about 10^{11} cm^{-2} .

Similar to bulk impurities, an interface trap is considered a donor if it is neutral and can become positively charged by donating (giving up) an electron. So, the donor state usually exists in the lower half of the bandgap, as shown in Fig. 11. Because a negative voltage is applied, the Fermi level moves down with respect to the interface-trap levels and the interface traps become positively charged. An acceptor interface trap is neutral and becomes negatively charged by accepting an electron. So, the acceptor state usually exists in the upper half of the bandgap also shown in Fig. 11.

Figure 11 shows an interface-trap system consisting of both acceptor states and donor states, in which the states above a neutral level E_0 are of the acceptor type and those below E_0 are of the donor type. To calculate the trapped charge, we can assume that at room temperature, the occupancy takes on the value of 0 and 1 above and below E_F . With these assumptions, the interface-trapped charges Q_{it} can now be easily calculated by:

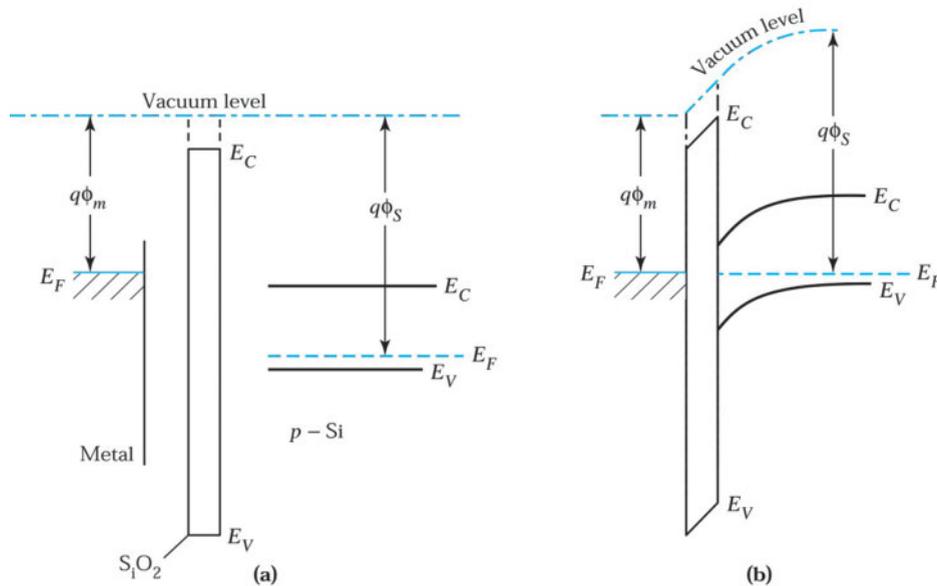


Fig. 9 (a) Energy band diagram of an isolated metal and an isolated semiconductor with an oxide layer between them. (b) Energy band diagram of an MOS capacitor in thermal equilibrium.

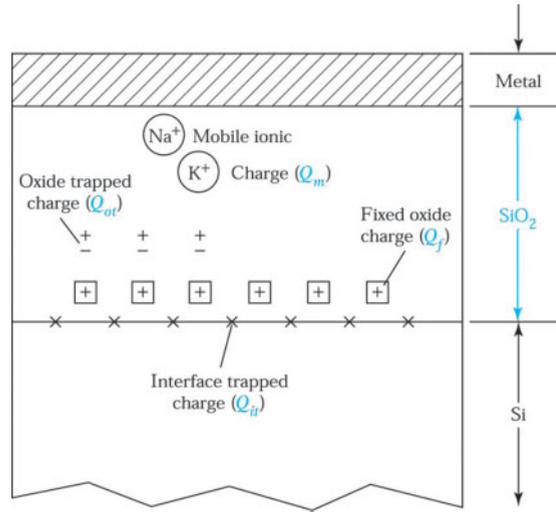


Fig. 10 Terminology for the charges associated with thermally oxidized silicon.³

$$\begin{aligned}
 Q_{it} &= -q \int_{E_0}^{E_F} D_{it} dE & E_F \text{ above } E_0, \\
 &= +q \int_{E_F}^{E_0} D_{it} dE & E_F \text{ below } E_0,
 \end{aligned}
 \tag{19}$$

where D_{it} is the interface trap density and Q_{it} is the effective net charges per unit area (i.e., C/cm²). The interface-trap levels are distributed across the energy bandgap and the interface trap density is given by:

$$D_{it} = \frac{1}{q} \frac{dQ_{it}}{dE} \quad \text{Number of traps / cm}^2 \text{ - eV.}
 \tag{20}$$

This is the method used to determine D_{it} experimentally from the change of Q_{it} in response to the change of E_F or surface potential ψ_s . On the other hand, Eq. 20 cannot distinguish whether the interface traps are of donor type or acceptor type but only determine the magnitude of D_{it} .

When a voltage is applied, the Fermi level moves up or down with respect to the interface-trap levels and a change of charge in the interface traps occurs. This change of charge affects the MOS capacitance and alters the ideal MOS curve.

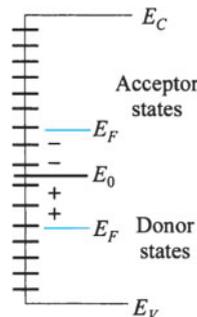


Fig. 11 Any interface-trap system consisting of both acceptor states and donor states can be interpreted by an equivalent distribution with a neutral level E_0 above which the states are of the acceptor type and below which of the donor type. When E_F is above (below) E_0 , the net charge is $-$ ($+$).

The fixed charge Q_f is located within approximately 3 nm of the SiO₂-Si interface. This charge is fixed and cannot be charged or discharged over a wide variation in surface potential ψ_s . Generally, Q_f is positive and depends on oxidation and annealing conditions and on silicon orientation. It has been suggested that when the oxidation is stopped, some ionic silicon is left near the interface. These ions, along with uncompleted silicon bonds (e.g., Si-Si or Si-O bonds) at the surface, may result in the positive interface charge Q_f . Q_f can be regarded as a charge sheet located at the SiO₂-Si interface. Typical fixed-oxide charge densities for a carefully treated SiO₂-Si interface system are about 10^{10} cm⁻² for a $\langle 100 \rangle$ surface and about 5×10^{10} cm⁻² for a $\langle 111 \rangle$ surface. Because of the lower values of Q_{it} and Q_f , the $\langle 100 \rangle$ orientation is preferred for silicon MOSFETs.

The oxide-trapped charges Q_{ot} are associated with defects in the silicon dioxide. These charges can be created, for example, by X-ray radiation or high-energy electron bombardment. The traps are distributed inside the oxide layer. Most of process-related Q_{ot} can be removed by low-temperature annealing.

The mobile ionic charges Q_m , such as sodium or other alkali ions, are mobile within the oxide under high-temperature (e.g., $>100^\circ\text{C}$) and high electric-field operations. Trace contamination by alkali metal ions may cause stability problems in semiconductor devices operated under high-bias and high-temperature conditions. Under these conditions mobile ionic charges can move back and forth through the oxide layer and cause shifts of the C - V curves along the voltage axis. Therefore, special attention must be paid to eliminate mobile ions in device fabrication.

The charges are the effective net charges per unit area in C/cm². We now evaluate the influence of these charges on the flat-band voltage. Consider a positive sheet charge per unit area, Q_o , within the oxide, as shown in Fig. 12. This positive sheet charge will induce negative charges partly in the metal and partly in the semiconductor, as shown in the upper part of Fig. 12a. The resulting field distribution, obtained from integrating Poisson's equation once, is shown in the lower part of Fig. 12a, where we have assumed that there is no work function difference, or $q\phi_{ms} = 0$.

To reach the flat-band condition (i.e., no charge induced in the semiconductor), we must apply a negative voltage to the metal, as shown in Fig. 12b. As the negative voltage increases, more negative charges are put on the metal and thereby the electric-field distribution shifts downward until the electric field at the semiconductor surface is zero. Under this condition, the area contained under the electric-field distribution corresponds to the flat-band voltage V_{FB} :

$$V_{FB} = \epsilon_o' x_o \left(\frac{Q_o}{\epsilon_{ox}} x_o - \frac{Q_o}{C_o} \frac{x_o}{d} \right) \quad (20)$$

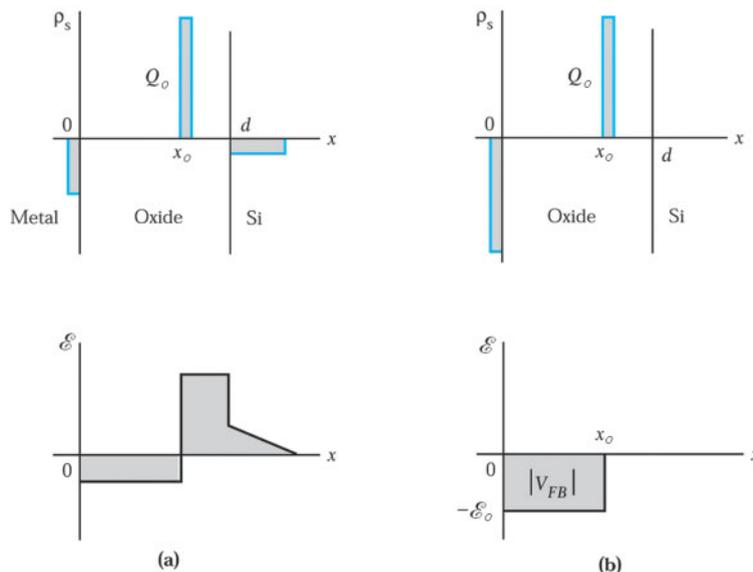


Fig. 12 Effect of a sheet charge within the oxide.² (a) Condition for $V_G = 0$. (b) Flat-band condition.

The flat-band voltage is thus dependent on both the density of the sheet charge Q_o and its location x_o within the oxide. When the sheet charge is located very close to the metal that is, if $x_o = 0$ it will induce no charges in the silicon and therefore have no effect on the flat-band voltage. On the other hand, when Q_o is located very close to the semiconductor $x_o = d$ such as the fixed-oxide charge Q_f , it will exert its maximum influence and give rise to a flat-band voltage

$$V_{FB} = -\frac{Q_o}{C_o} \frac{d}{d} = -\frac{Q_o}{C_o}. \quad (22)$$

For the more general case of an arbitrary space charge distribution within the oxide, the flat-band voltage is given by

$$V_{FB} = \frac{-1}{C_o} \left[\frac{1}{d} \int_0^d x \rho(x) dx \right], \quad (23)$$

where $\rho(x)$ is the volume charge density in the oxide. Once we know $\rho_{ot}(x)$, the volume charge density for oxide-trapped charges, and $\rho_m(x)$, the volume charge density for mobile ionic charges, we can obtain Q_{ot} and Q_m and their corresponding contribution to the flat-band voltage:

$$Q_{ot} \equiv \frac{1}{d} \int_0^d x \rho_{ot}(x) dx, \quad (24a)$$

$$Q_m \equiv \frac{1}{d} \int_0^d x \rho_m(x) dx. \quad (24b)$$

If the value of the work function difference $q\phi_{ms}$ is not zero and if the value of the interface-trapped charges is negligible, the experimental capacitance-voltage curve will be shifted from the ideal theoretical curve by an amount

$$V_{FB} = \phi_{ms} - \frac{(Q_f + Q_m + Q_{ot})}{C_o}. \quad (25)$$

The curve in Fig. 13a shows the C - V characteristics of an ideal MOS capacitor. Due to nonzero ϕ_{ms} , Q_f , Q_m , or Q_{ot} , the C - V curve will be shifted by an amount given by Eq. 25. The parallel shift of the C - V curve is illustrated in Fig. 13b. If, in addition, there are large amounts of interface-trapped charges, the charges in the interface traps will vary with the surface potential. The C - V curve will be displaced by an amount that itself changes with the surface potential. Therefore, Fig. 13c is distorted as well as shifted because of interface-trapped charges.

► EXAMPLE 3

Calculate the flat-band voltage for an n^+ -polysilicon-SiO₂-Si capacitor having $N_A = 10^{17} \text{ cm}^{-3}$ and $d = 5 \text{ nm}$. Assume that Q_i and Q_m are negligible in the oxide, and Q_f/q is $5 \times 10^{11} \text{ cm}^{-2}$.

SOLUTION From Fig. 8, ϕ_{ms} is -0.98 V for n^+ polysilicon (p -Si) system with $N_A = 10^{17} \text{ cm}^{-3}$. C_o is obtained from Ex. 2.

$$V_{FB} = \phi_{ms} - \frac{(Q_f + Q_m + Q_{ot})}{C_o}$$

$$= -0.98 - \frac{(1.6 \times 10^{-19} \times 5 \times 10^{11})}{6.9 \times 10^{-7}} = -1.10 \text{ V.}$$

► **EXAMPLE 4**

Assume that the volume charge density, $\rho_{ot}(x)$, for oxide-trapped charge Q_{ot} in an oxide layer has a triangular distribution. The distribution is described by the function $(10^{18} - 5 \times 10^{23} \times x) \text{ cm}^{-3}$, where x is the distance from the location to the metal-oxide interface. The thickness of the oxide layer is 20 nm. Find the change in the flat-band voltage due to Q_{ot} .

SOLUTION From Eqs. 23 and 24a,

$$\begin{aligned} \Delta V_{FB} &= \frac{Q_{ot}}{C_o} = \frac{d}{\epsilon_{ox}} \frac{1}{d} \int_0^{2 \times 10^{-6}} x \rho_{ot}(x) dx \\ &= \frac{1.6 \times 10^{-19}}{3.9 \times 8.85 \times 10^{-14}} \left[\frac{1}{2} \times 10^{18} \times (2 \times 10^{-6})^2 - \frac{1}{3} \times 5 \times 10^{23} \times (2 \times 10^{-6})^3 \right] \\ &= \frac{1.6 \times 10^{-19} \times (2 \times 10^6 - 1.33 \times 10^6)}{3.45 \times 10^{-13}} \\ &= 0.31 \text{ V.} \end{aligned}$$

► **5.3 CARRIER TRANSPORT IN MOS CAPACITORS**

In an ideal MOS capacitor, the conductance of the insulating film is assumed to be zero. Real insulators, however, show some degree of carrier conduction when the electric field or temperature is high.

5.3.1 Basic Conduction Processes in Insulators

Tunneling is the conduction mechanism through insulators under high fields. The tunneling emission is a result of quantum mechanics by which the electron wave function can penetrate a potential barrier. From Section 2.6 of Chapter 2, the tunneling current is proportional to the transmission coefficient $\exp(-2\beta d)$, where d is the insulator thickness and $\beta \sim (qV_0 - E)^{1/2} \sim \{[E_1 + (E_2 - qV)]/2\}^{1/2}$. The term $[E_1 + (E_2 - qV)]/2$ is the average potential barrier height, where E_1 and E_2 are the barrier heights shown in Fig. 14a and V is

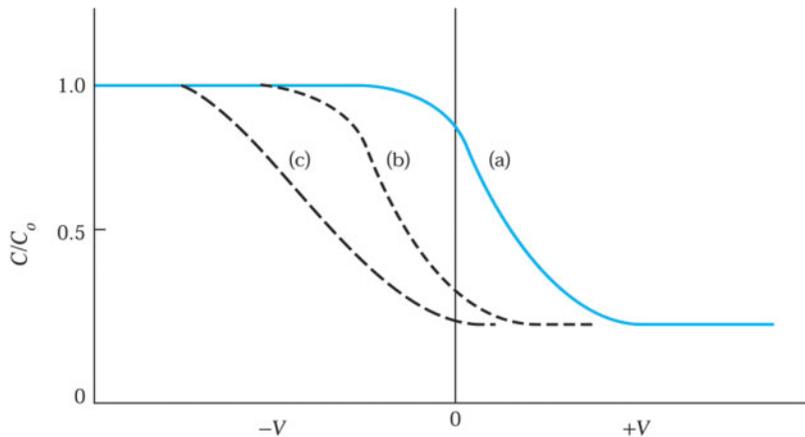


Fig. 13 (a) The C - V characteristics of an ideal MOS capacitor. (b) Parallel shift along the voltage axis due to positive fixed-oxide charges. (c) Nonparallel shift along the voltage axis due to interface traps.

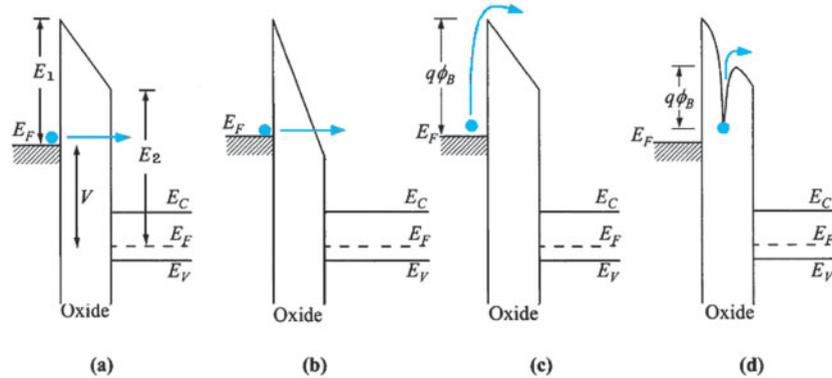


Fig. 14 Energy-band diagrams showing conduction mechanisms of (a) direct tunneling, (b) Fowler-Nordheim tunneling, (c) thermionic emission, and (d) Frenkel-Poole emission.

the applied voltage. When V increases, β will decrease and the transmission coefficient as well as the tunneling current will increase. Therefore, the current is dependent on the applied voltage but is independent of temperature. Figure 14a is for the direct tunneling, i.e., tunneling through the complete width of the insulator. Figure 14b is for the Fowler-Nordheim tunneling in which carrier tunneling through only a partial width of the barrier. In this case, both the average potential barrier height and the tunneling distance are reduced from those of direct tunneling.

The thermionic emission (Schottky emission) process is from the carrier transport of electrons with energies sufficient to overcome the metal-insulator barrier or the insulator-semiconductor barrier shown in Fig. 14c. From Section 2.5 of Chapter 2, the thermionic emission current is proportional to the electron density with energies above the barrier height, i.e. $q\chi$ for a vacuum-semiconductor interface or $q\phi_B$ for a metal-insulator interface. Therefore, for the MOS capacitor the current is proportional to $\exp(q\phi_B/kT)$. It increases exponentially with decreasing barrier height and increasing temperature.

The Frenkel-Poole emission, shown in Fig. 14d, is due to the emission of trapped electrons into the conduction band through thermal excitation. The emission is similar to that of the Schottky emission. The barrier height, however, is the depth of the trap potential well.

At low voltage and high temperature, current is carried by thermally excited electrons hopping from one isolated state to the next. This mechanism yields an ohmic characteristic exponentially dependent on temperature.

The ionic conduction is similar to a diffusion process. Generally, the dc ionic conductivity decreases during the time the electric field is applied because ions cannot be readily injected into or extracted from the insulator. After an initial current flow, positive and negative space charges will build up near the metal-insulator and the semiconductor-insulator interfaces, causing a distortion in the potential distribution. When the applied field is removed, large internal fields remain that cause some, but not all, ions to flow back toward their equilibrium position. This will result in an I - V hysteresis.

The space-charge-limited current results from carriers injected into a lightly doped semiconductor or an insulator, where no compensating charge is present. The current is proportional to the square of the applied voltage.

For a given insulator, each conduction process may dominate in a certain temperature and voltage range. Figure 15 shows plots of current density versus $1/T$ for three different insulators, Si_3N_4 , Al_2O_3 , and SiO_2 ⁴. The conduction here can be divided into three temperature ranges. At high temperatures (and high fields), the current J_1 is due to Frenkel-Poole emission. At intermediate temperature, the current J_3 is ohmic in nature. At low temperatures, the conduction is tunneling limited and the current J_2 is temperature insensitive. One can also observe that the tunneling current strongly depends on the barrier height, which is related to the energy gap of the insulators (Si_3N_4 (4.7 eV) < Al_2O_3 (8.8 eV) < SiO_2 (9 eV)). The larger the energy gap, the lower the current. The current in SiO_2 is three orders of magnitude lower than that in Si_3N_4 .

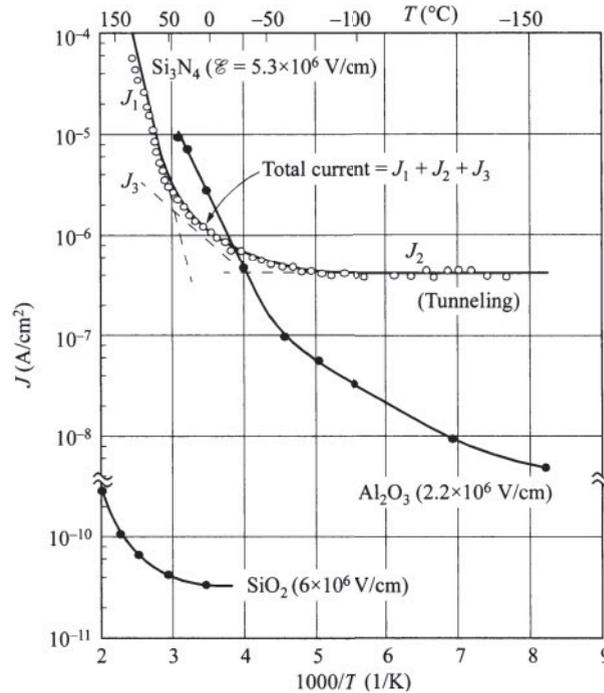


Fig. 15 Current density versus $1/T$ for Si_3N_4 , Al_2O_3 , and SiO_2 films.

5.3.2 Dielectric Breakdown

Microscopically, the percolation theory shown in Fig. 16 is used to explain breakdown.⁴ Under a large bias, some current will conduct through the insulator, most commonly a tunneling current. When energetic carriers move through the insulator, defects are generated randomly in the bulk of the dielectric film. When defects are dense enough to form a continuous chain connecting the gate to the semiconductor, a conduction path is created and catastrophic breakdown occurs.

A measure to quantify reliability is time to breakdown, t_{BD} , which is the total stress time until breakdown occurs. An example for t_{BD} versus oxide field for different oxide thickness is shown in Fig. 17.⁴ A few key points can be noticed in the figure. First, t_{BD} is a function of bias. Even for a small bias, eventually the oxide will break down, taking a very long time. Conversely, a large field can be sustained for a very short time without breaking down. In addition, the breakdown field decreases as the oxide becomes thicker. This is because for a given electric field, a higher voltage is required for a thicker film. A higher voltage provides higher energy for the carriers, causing more damages to the oxide and reducing t_{BD} .

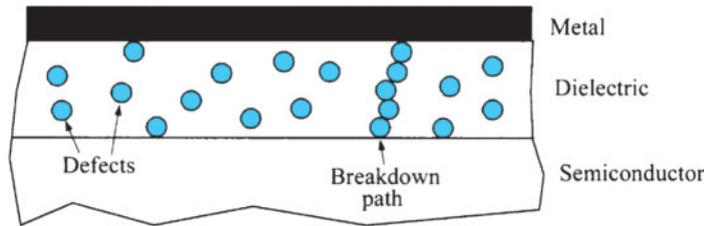


Fig. 16 Percolation theory: breakdown occurs when random defects form a chain between the gate and the semiconductor.

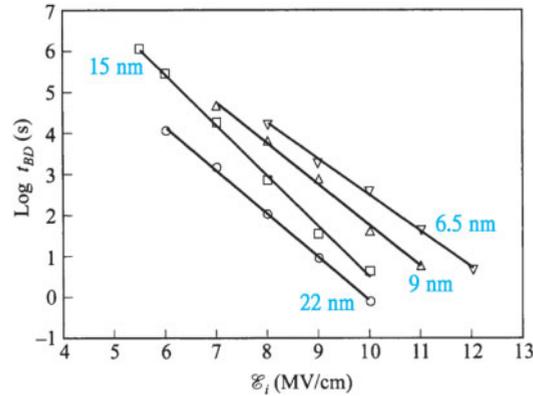


Fig. 17 Time to breakdown t_{BD} vs. oxide field, for different oxide thicknesses.

► 5.4 CHARGE-COUPLED DEVICES (CCD)

A schematic view of a CCD is shown⁵ in Fig. 18. The basic device consists of a closely spaced array of MOS capacitors on a continuous insulator (oxide) layer that covers the semiconductor substrate. A CCD can perform a wide range of electronic functions, including image sensing and signal processing. The operating principle of the CCD involves the charge storage and transfer actions controlled by the gate electrodes.

Figure 18a shows a CCD to which sufficiently large positive bias pulses have been applied to all the electrodes to produce surface depletion. A slightly higher bias has been applied to the central electrode so that the center MOS structure is under greater depletion and a potential well is formed there; i.e., the potential distribution is shaped like a well because of the larger depletion-layer width under the central electrode. If minority carriers (electrons) are introduced, they will be collected in the potential well. If the potential of the right-hand electrode is increased to exceed that of the central electrode, we obtain the potential distribution shown in Fig. 18b. In this case, the minority carriers will be transferred from the central electrode to the right-hand electrode. Subsequently, the potential on the electrodes can be readjusted so that the quiescent storage site is located at the right-hand electrode. By continuing this process, we can transfer the carriers successively along a linear array.

CCD Shift Register

Figure 19 shows more details about the basic principle of charge transfer in a three-phase, n -channel CCD array. The electrodes are connected to the ϕ_1 , ϕ_2 , and ϕ_3 clock lines. Figure 19b shows the clock waveforms and Fig. 19c illustrates the corresponding potential wells and charge distributions.

At $t = t_1$, clock line ϕ_1 is at a high voltage and ϕ_2 and ϕ_3 are at low voltages. The potential wells under ϕ_1 will be deeper than the others. We assume that there is a signal charge at the first ϕ_1 electrode. At $t = t_2$, both ϕ_1 and ϕ_2 have high bias as charge starts to transfer. At $t = t_3$, the voltage at ϕ_1 is returning to the low value while ϕ_2 electrodes are still held at high voltage. The electrons stored under ϕ_1 are being emptied in this period. At $t = t_4$, the charge transfer is complete and the original charge packet is now stored under the first ϕ_2 electrode. This process will be repeated and the charge packet continues to shift to the right. CCDs can be operated with two, three, or four phases, with different design structures. Multiple electrode structures and clocking schemes have been proposed and implemented.⁶

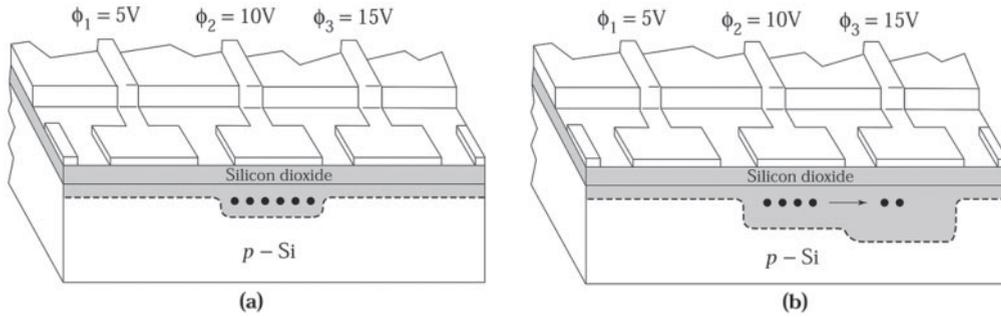


Fig. 18 Cross section of a three-phase charge-coupled device.⁵ (a) High voltage on ϕ_2 . (b) ϕ_3 pulsed to a higher voltage for charge transfer.

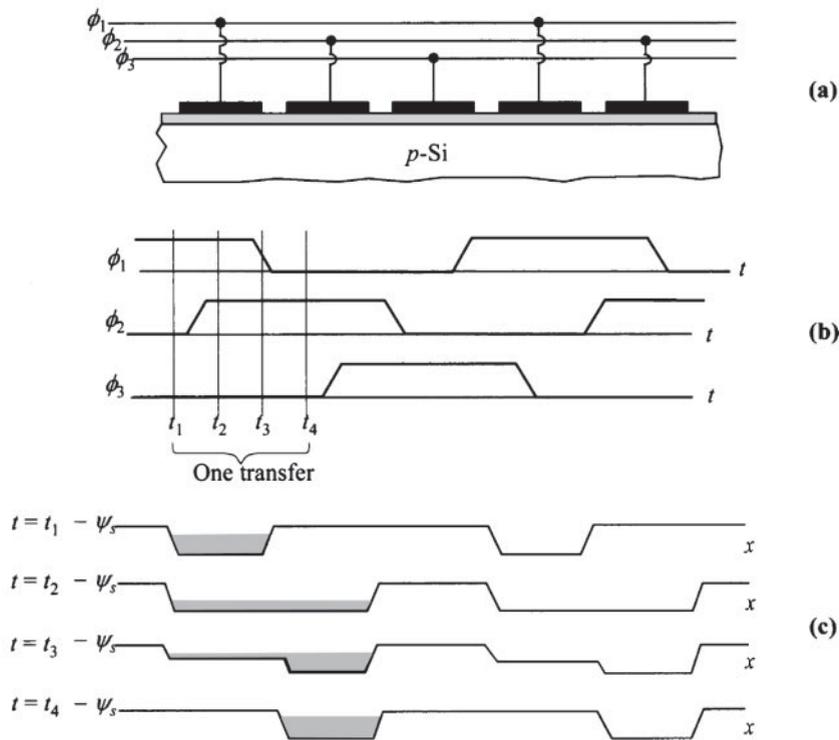


Fig. 19 Illustration of CCD charge transfer. (a) Application of three-phase gate bias. (b) Clock waveforms. (c) Surface potential (and charge) vs. distance at different times.⁶

CCD Image Sensor

For analog and memory devices, the charge packets are introduced by injection from a $p-n$ junction in the vicinity of the CCD. For optical imaging applications, the charge packets are formed as a result of electron-hole pair generation caused by incident light.

When CCD used in imaging array systems such as a camera or video recorder, CCD image sensors must be spaced closely to one another in a chain and function as shift registers to transport the signals. The structure of the surface-channel CCD image sensor is similar to that of the CCD shift register, with the exception that the gates are semitransparent to let light pass through. Common materials for the gates are metal, polysilicon, and silicide.

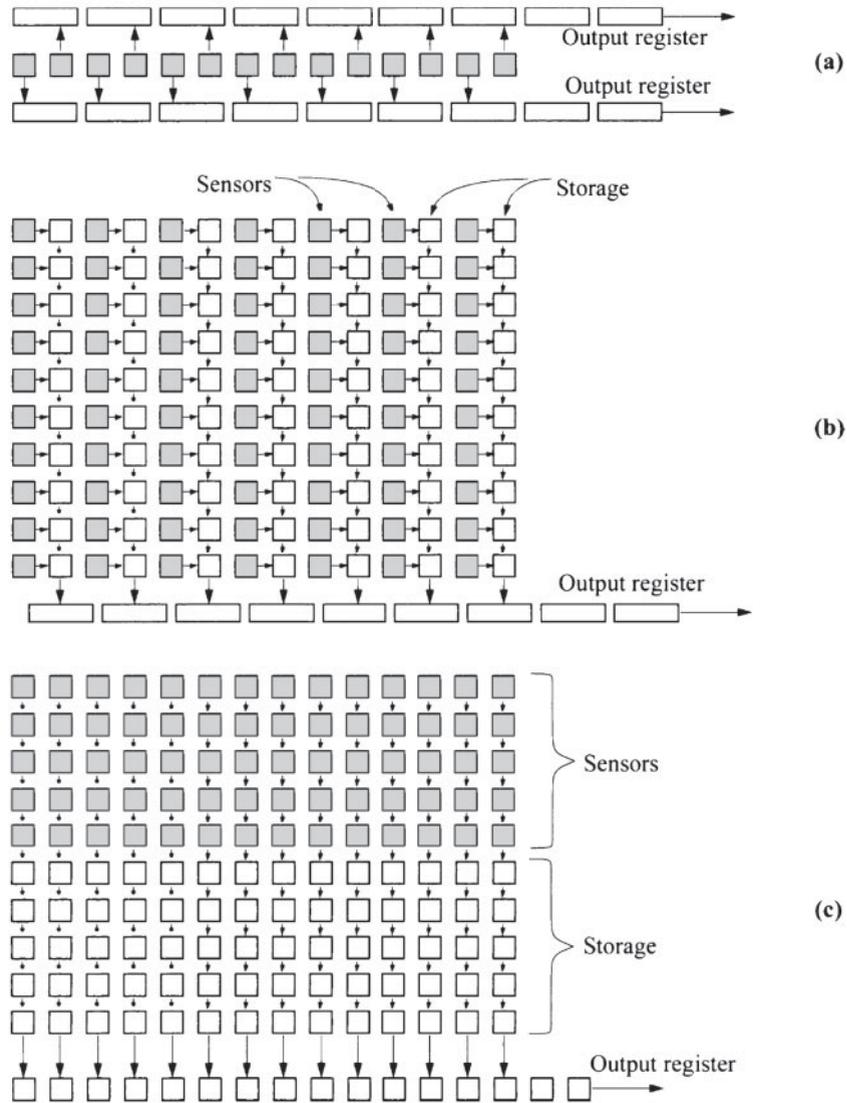


Fig. 20 Schematic layouts showing the readout mechanisms of (a) line imager with dual output registers, and area imagers with (b) interline transfer and (c) frame transfer. Gray pixels represent CCDs as photodetectors. The output register is usually clocked at higher frequency than the internal transfer.

Alternatively, the CCD can be illuminated from the back of the substrate to avoid light absorption by the gate. In this configuration, the semiconductor has to be thinned so that most of the light can be absorbed within the depletion region at the top surface.

Because the CCDs can also be used as a shift registers, there is great benefit to using CCDs as photodetectors in an imaging-array system since the signals can be brought out sequentially to a single node, without complicated x - y addressing to each pixel. The photogenerated carriers are integrated during light exposure, and the signal is stored in the form of a charge packet, to be transported and detected later. The detection mode of the integrated charge over a long period of time enables detection of weaker signals. In addition, the CCDs have the advantages of low dark current, low noise, low-voltage operation, good linearity, and good dynamic range. The structure is simple, compact, stable, and robust, and is compatible with MOS technology. These factors contribute to high yield, which makes the CCDs desirable in consumer products.

Different readout mechanisms for the line imager and the area imagers are shown in Fig. 20.⁴ A line imager with dual output registers has improved readout speed (Fig. 20a). Most common area imagers use either interline-transfer (Fig. 20b) or frame-transfer (Fig. 20c) readout architecture. In the former case, signals are transferred to the neighboring pixels and are subsequently passed along to the output register chain while the light-sensitive pixels start to collect a charge for the next data. In the frame-transfer scheme, signals are shifted to a storage area away from the sensing area. The advantage of this over the interline transfer is a more efficient light-sensing area, but there is more image smear since CCDs continue to receive light as signal charges are passed through them. For both interline transfer and frame transfer, all columns advance their charge signals to the horizontal output register simultaneously, and the output register carries these signals out at a much higher clocking rate.

► 5.5 MOSFET FUNDAMENTALS

The **MOSFET** has many acronyms, including **IGFET** (insulating-gate field-effect transistor), **MISFET** (metal-insulator-semiconductor field-effect transistor) and **MOST** (metal-oxide-semiconductor transistor). A perspective view of an n -channel **MOSFET** is shown in Fig. 21. It is a four-terminal device consisting of a p -type semiconductor substrate in which two n^+ regions, the source and drain, are formed. The metal plate on the oxide is called the gate. Heavily doped polysilicon or a combination of a silicide such as WSi_2 and polysilicon can be used as the gate electrode. The fourth terminal is an ohmic contact to the substrate. The basic device parameters are the channel length L , which is the distance between the two metallurgical n^+ - p junctions, the channel width Z , the oxide thickness d , the junction depth r_j , and the substrate doping N_A .⁸ Note that the central section of the device corresponds to the MOS capacitor discussed in Section 5.1.

The first MOSFET was fabricated in 1960 using a thermally oxidized silicon substrate.⁷ The device had a channel length of 20 μm and a gate oxide thickness of 100 nm.* Although present-day MOSFETs have been scaled down considerably, the silicon and thermally grown silicon dioxide used in the first MOSFET remains the most important combination.⁸ Most of the results in this section are obtained from the Si-SiO₂ system.

⁸For p -channel MOSFETs, doping types in substrate and source/drain regions become n and p^+ , respectively.

* A photograph of the first MOSFET is shown in Fig. 4 of Chapter 0.

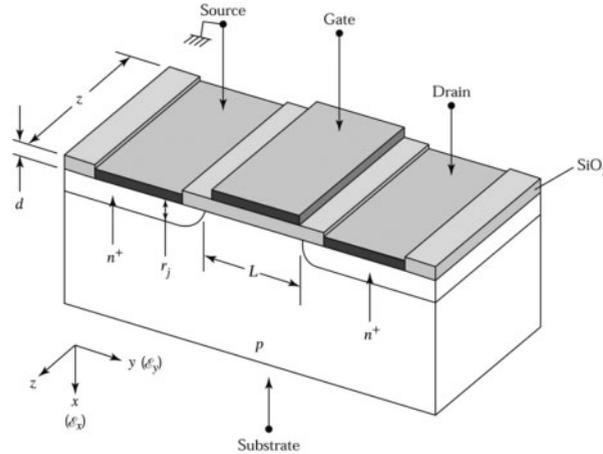


Fig. 21 Perspective view of a metal-oxide-semiconductor field-effect transistor (MOSFET).

5.5.1 Basic Characteristics

The source contact is used as the voltage reference throughout this section. When no voltage is applied to the gate, the source-to-drain electrodes correspond to two p - n junctions connected back to back. The only current that can flow from the source to drain is the reverse-leakage current.[†] When we apply a sufficiently large positive bias to the gate, the MOS structure is inverted so that a surface inversion layer (or channel) is formed between the two n^+ -regions. The source and drain are then connected by a conducting surface n -channel through which a large current can flow. The conductance of this channel can be modulated by varying the gate voltage. The substrate contact can be at the reference voltage or is reverse biased with respect to the source; the substrate bias voltage will also affect the channel conductance.

Linear and Saturation Regions

We now present a qualitative discussion of MOSFET operation. Let us consider that a voltage is applied to the gate, causing an inversion at the semiconductor surface (Fig. 22). If a small drain voltage is applied, electrons will flow from the source to the drain (the corresponding current will flow from drain to source) through the conducting channel. Thus, the channel acts as a resistor, and the drain current I_D is proportional to the drain voltage. This is the *linear region*, as indicated by the constant-resistance line in the right-hand diagram of Fig. 22a.

When the drain voltage increases, eventually it reaches V_{Dsat} , at which the thickness of the inversion layer x_i near $y = L$ is reduced to zero; this is called the pinch-off point, P (Fig. 22b). Beyond the pinch-off point, the drain current remains essentially the same, because for $V_D > V_{Dsat}$, at point P the voltage V_{Dsat} remains the same. Thus, the number of carriers arriving at point P from the source or the current flowing from the drain to the source remains the same. This is the *saturation region*, since I_D is a constant regardless of an increase in the drain voltage. The major change is the decrease of L to the value L' shown in Fig. 22c. Carrier injection from P into the drain depletion region is similar to that of carrier injection from an emitter-base junction to the base-collector depletion region of a bipolar transistor.

We now derive the basic MOSFET characteristics under the following ideal conditions. (a) The gate structure corresponds to an ideal MOS capacitor, as defined in Section 5.1, that is, there are no interface traps, fixed-oxide charges, or work function differences. (b) Only drift current is considered. (c) Carrier mobility in the inversion layer is constant. (d) Doping in the channel is uniform. (e) Reverse-leakage current is negligibly small. (f) The transverse field created by the gate voltage (\mathcal{E}_x in the x -direction, shown in Fig. 21, which is perpendicular to the current flow) in the channel is much larger than the longitudinal field created by the drain voltage (\mathcal{E}_y in the y -direction,

[†] This is true for the n -channel, normally off MOSFET. Other types of MOSFET are discussed in Section 5.5.2.

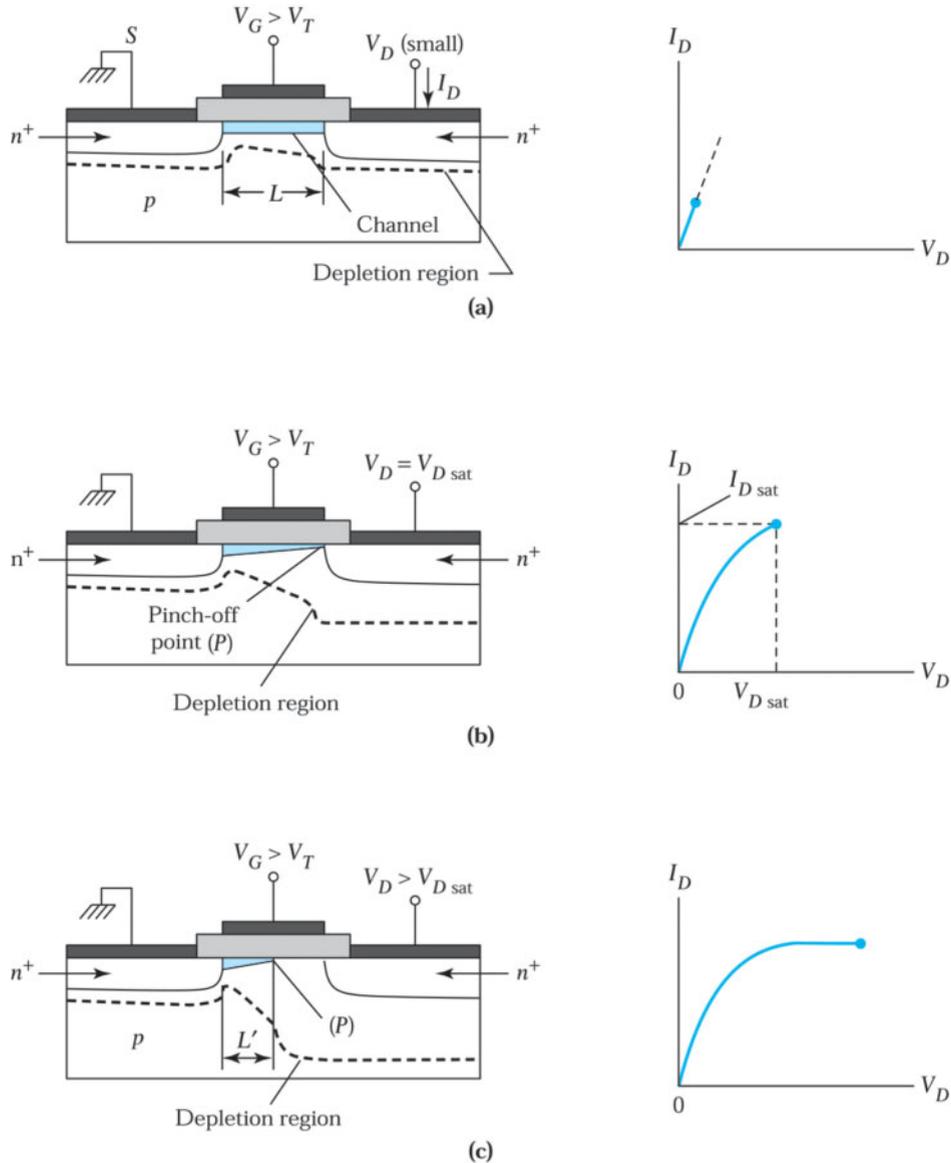


Fig. 22 Operations of the MOSFET and output I - V characteristics. (a) Low drain voltage. (b) Onset of saturation. Point P indicates the pinch-off point. (c) Beyond saturation.

which is parallel to the current flow). The last condition is called the gradual-channel approximation and generally is valid for long-channel MOSFETs. Under this approximation, the charges contained in the surface depletion region of the substrate are induced solely from the field created by the gate voltage.

Figure 23a shows the MOSFET operated in the linear region. Under the above ideal conditions, the total charge induced in the semiconductor per unit area, Q_s , at a distance y from the source is shown in Fig. 23b, which is an enlarged central section of Fig. 23a. Q_s is given from Eqs. 13 and 14 by

$$Q_s(y) = -[V_G - \psi_s(y)]C_o, \tag{26}$$

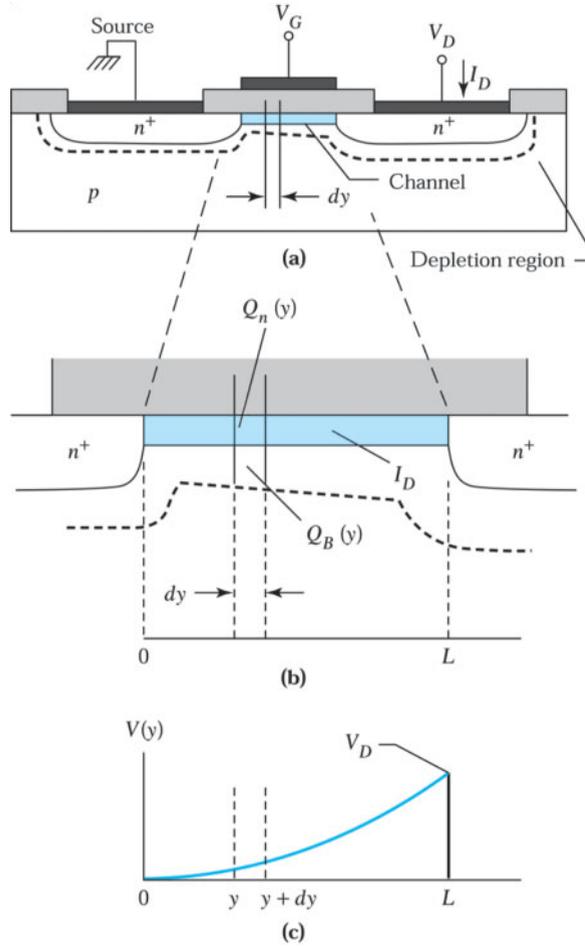


Fig. 23 (a) MOSFET operated in the linear region. (b) Enlarged view of the channel. (c) Drain voltage drop along the channel.

where $\psi_s(y)$ is the surface potential at y and $C_o = \epsilon_{ox}/d$ is the gate capacitance per unit area. Since Q_s is the sum of the charge in the inversion layer per unit area, Q_n , and the charge in surface depletion region per unit area, Q_{sc} , we can obtain Q_n as

$$\begin{aligned} Q_n(y) &= Q_s(y) - Q_{sc}(y), \\ &= -[V_G - \psi_s(y)]C_o - Q_{sc}(y). \end{aligned} \quad (27)$$

The surface potential $\psi_s(y)$ at inversion can be approximated by $2\psi_B + V(y)$, where $V(y)$, as shown in Fig. 23c, is the reverse bias between the point y and the source electrode (which is assumed to be grounded). The charge within the surface depletion region $Q_{sc}(y)$ was given previously as

$$Q_{sc}(y) = -qN_A W_m \cong -\sqrt{2\epsilon_s q N_A [2\psi_B + V(y)]}. \quad (28)$$

Substituting Eq. 28 in Eq. 27 yields

$$Q_n(y) \cong -[V_G - V(y) - 2\psi_B]C_o + \sqrt{2\epsilon_s q N_A [2\psi_B + V(y)]}. \quad (29)$$

The conductivity of the channel at position y can be approximated by

$$\sigma(x) = qn(x)\mu_n(x). \quad (30)$$

For a constant mobility, the channel conductance is then given by

$$g = \frac{Z}{L} \int_0^{x_i} \sigma(x) dx = \frac{Z\mu_n}{L} \int_0^{x_i} qn(x) dx. \quad (31)$$

The integral $\int_0^{x_i} qn(x) dx$ corresponds to the total charge per unit area in the inversion layer and is therefore equal to $|Q_n|$, or

$$g = \frac{Z\mu_n}{L} |Q_n|. \quad (32)$$

The channel resistance of an elemental section dy (Fig. 23b) is

$$dR = \frac{dy}{gL} = \frac{dy}{Z\mu_n |Q_n(y)|}, \quad (33)$$

and the voltage drop across the elemental section is

$$dV = I_D dR = \frac{I_D dy}{Z\mu_n |Q_n(y)|}, \quad (34)$$

where I_D is the drain current, which is independent of y . Substituting Eq. 29 into Eq. 34 and integrating from the source ($y = 0, V = 0$) to the drain ($y = L, V = V_D$) yield

$$I_D \approx \frac{Z}{L} \mu_n C_o \left\{ \left(V_G - 2\psi_B - \frac{V_D}{2} \right) V_D - \frac{2}{3} \frac{\sqrt{2\varepsilon_s q N_A}}{C_o} \left[(V_D + 2\psi_B)^{3/2} - (2\psi_B)^{3/2} \right] \right\} \quad (35)$$

Figure 24 shows the current-voltage characteristics of an idealized MOSFET based on Eq. 35. For a given V_G , the drain current first increases linearly with drain voltage (the linear region), then gradually levels off, approaching a saturated value (the saturation region). The dashed line indicates the locus of the drain voltage (V_{Dsat}) at which the current reaches a maximum value.

We now consider the linear and saturation regions. For small V_D , Eq. 35 reduces to

$$I_D \cong \frac{Z}{L} \mu_n C_o \left(V_G - V_T - \frac{V_D}{2} \right) V_D \quad \text{for } V_D < (V_G - V_T). \quad (36)$$

For very small V_D , Eq. 35 reduces to

$$I_D \cong \frac{Z}{L} \mu_n C_o (V_G - V_T) V_D \quad \text{for } V_D \ll (V_G - V_T), \quad (36a)$$

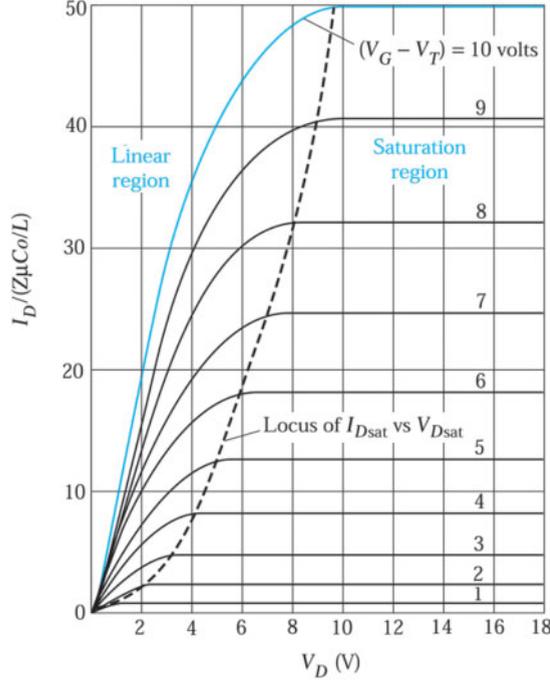


Fig. 24 Idealized drain characteristics of MOSFET. For $V_D \geq V_{Dsat}$, the drain current remains constant.

where V_T is the threshold voltage given previously in Eq. 17:

$$V_T = \frac{\sqrt{2\epsilon_s q N_A (2\psi_B)}}{C_o} + 2\psi_B. \quad (37)$$

By plotting I_D versus V_G (for a given small V_D), the threshold voltage can be deduced from the linearly extrapolated value at the V_G axis. In the linear region, Eq. 36, the channel conductance g_D and the transconductance g_m are given as

$$g_D \equiv \left. \frac{\partial I_D}{\partial V_D} \right|_{V_G \text{ constant}} \cong \frac{Z}{L} \mu_n C_o (V_G - V_T - V_D), \quad (38)$$

$$g_m \equiv \left. \frac{\partial I_D}{\partial V_G} \right|_{V_D \text{ constant}} \cong \frac{Z}{L} \mu_n C_o V_D. \quad (39)$$

When the drain voltage is increased to a point that the charge $Q_n(y)$ in the inversion layer at $y = L$ becomes zero (pinch-off), the number of mobile electrons at the drain are reduced drastically. The drain voltage and the drain current at this point are designated as V_{Dsat} and I_{Dsat} , respectively. For drain voltages larger than V_{Dsat} , we have the saturation region. We can obtain the value of V_{Dsat} from Eq. 29 under the condition $Q_n(L) = 0$:

$$V_{Dsat} \cong V_G - 2\psi_B + K^2 \left(1 - \sqrt{1 + \frac{2V_G}{K^2}} \right), \quad (40)$$

where $K \equiv \frac{\sqrt{\epsilon_s q N_A}}{C_o}$. The saturation current can be obtained by substituting Eq. 40 into Eq. 35:

$$I_{Dsat} \cong \left(\frac{Z \mu_n C_o}{2L} \right) (V_G - V_T)^2. \quad (41)$$

The threshold voltage V_T in the saturation region for low substrate doping and thin oxide layers is the same as that from Eq. 37. At higher doping levels, V_T becomes V_G dependent.

For an idealized MOSFET in the saturation region, the channel conductance is zero, and the transconductance can be obtained from Eq. 41:

$$g_m \equiv \left. \frac{\partial I_D}{\partial V_G} \right|_{V_D \text{ constant}} = \frac{Z \mu_n \epsilon_{ox}}{dL} (V_G - V_T). \quad (42)$$

► EXAMPLE 5

For an n -channel n^+ -polysilicon-SiO₂-Si MOSFET with gate oxide = 8 nm, $N_A = 10^{17} \text{ cm}^{-3}$ and $V_G = 3\text{V}$, calculate V_{Dsat} .

SOLUTION

$$C_o = \frac{\epsilon_{ox}}{d} = \frac{3.9 \times 8.85 \times 10^{-14}}{8 \times 10^{-7}} = 4.32 \times 10^{-7} \text{ F/cm}^2$$

$$K = \frac{\sqrt{\epsilon_s q N_A}}{C_o} = \frac{\sqrt{11.9 \times 8.85 \times 10^{-14} \times 1.6 \times 10^{-19} \times 10^{17}}}{4.32 \times 10^{-7}} = 0.3$$

$2\psi_B = 0.84 \text{ V}$ from Ex. 2. Therefore, from Eq. 40,

$$\begin{aligned} V_{Dsat} &\cong V_G - 2\psi_B + K^2 \left(1 - \sqrt{1 + \frac{2V_G}{K^2}} \right) \\ &= 3 - 0.84 + (0.3)^2 \left[1 - \sqrt{1 + \frac{2 \times 3}{(0.3)^2}} \right] \\ &= 3 - 0.84 - 0.65 = 1.51 \text{ V}. \end{aligned}$$

The Subthreshold Region

When the gate voltage is below the threshold voltage and the semiconductor surface is only weakly inverted, the corresponding drain current is called the *subthreshold current*. The subthreshold region is particularly important when the MOSFET is used as a low-voltage, low-power device, such as a switch in digital logic and memory applications, because the subthreshold region describes how the switch turns on and off.

In the subthreshold region, the drain current is dominated by diffusion instead of drift and is derived in the same way as the collector current in a bipolar transistor with homogeneous base doping. If we consider the MOSFET as an n - p - n (source-substrate-drain) bipolar transistor (Fig. 23b), we have

$$I_D = -qAD_n \frac{\partial n}{\partial y} = -qAD_n \frac{n(0) - n(L)}{L}, \quad (43)$$

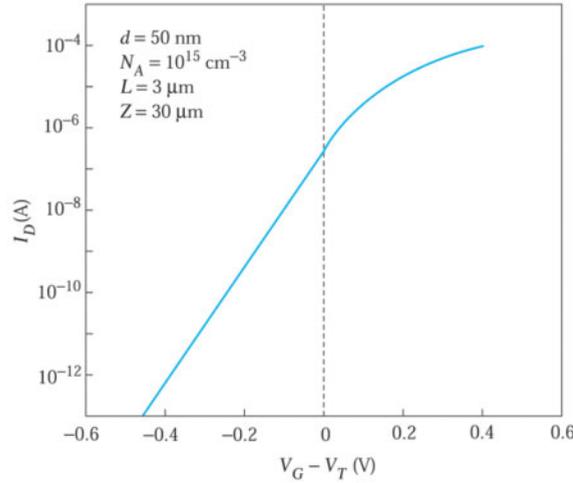


Fig. 25 Subthreshold characteristics of an MOSFET.

where A is the channel cross section of the current flow and $n(0)$ and $n(L)$ are the electron densities in the channel at the source and drain, respectively. The electron densities are given by Eq. 5a:

$$n(0) = n_i e^{q(\psi_s - \psi_B)/kT}, \quad (44a)$$

$$n(L) = n_i e^{q(\psi_s - \psi_B - V_D)/kT}, \quad (44b)$$

where ψ_s is the surface potential at the source. Substituting Eq. 44 into Eq. 43 gives

$$I_D = \frac{qAD_n n_i e^{-q\psi_B/kT}}{L} (1 - e^{-qV_D/kT}) e^{q\psi_s/kT}. \quad (45)$$

The surface potential ψ_s is approximately $V_G - V_T$. Therefore, the drain current will decrease exponentially when V_G becomes less than V_T :

$$I_D \sim e^{q(V_G - V_T)/kT}. \quad (46)$$

A typical measured curve for the subthreshold region is shown in Fig. 25. Note the exponential dependence of I_D on $(V_G - V_T)$ for $V_G < V_T$. An important parameter in this region is the *subthreshold swing*, S , which is defined as $\ln 10 [dV_G/d(\ln I_D)]$. The parameter quantifies how sharply the transistor is turned off by the gate voltage and is given by the gate-voltage change needed to induce a drain-current change of one order of magnitude. S is typically 70 ~ 100 mV/decade of drain current at room temperature. To reduce the subthreshold current to a negligible value, we must bias the MOSFET a half-volt or more below V_T .

5.5.2 Types of MOSFET

There are basically four types of MOSFETs, depending on the type of inversion layer. If, at zero gate bias, the channel conductance is very low and we must apply a positive voltage to the gate to form the n -channel, then the device is a normally off (enhancement) n -channel MOSFET. If an n -channel exists at zero bias and we must apply a negative voltage to the gate to deplete carriers in the channel to reduce the channel conductance, then the device is a normally on (depletion) n -channel MOSFET. Similarly, we have the p -channel normally off (enhancement) and normally on (depletion) MOSFETs.

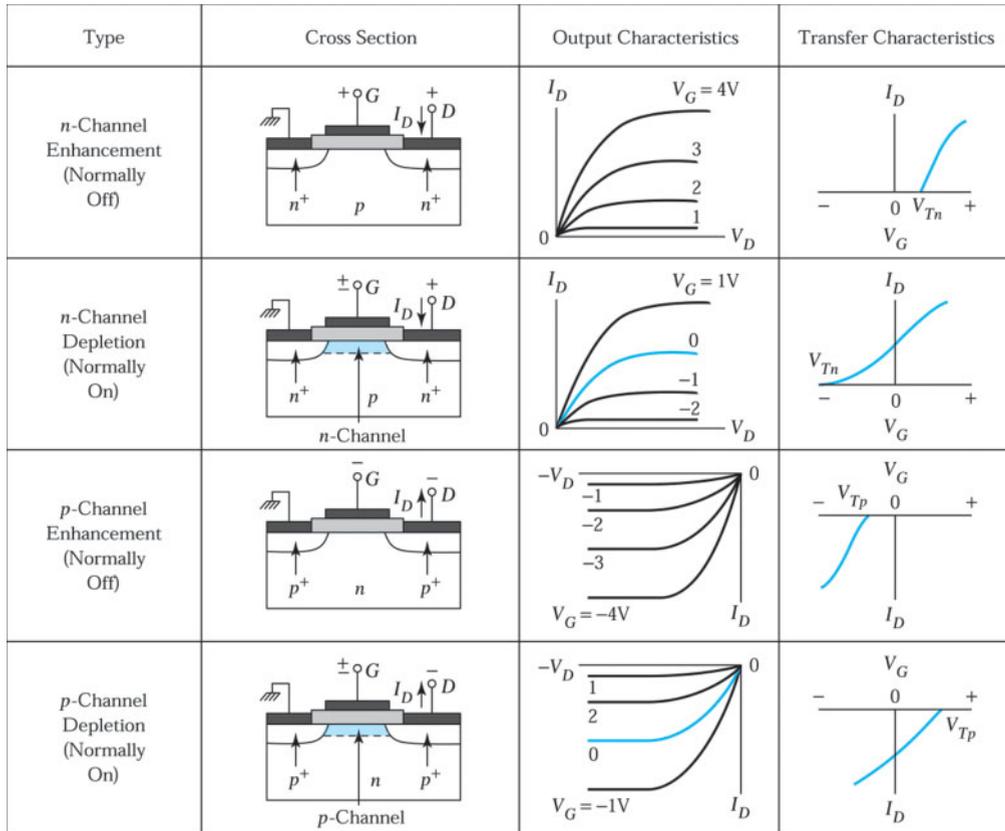


Fig. 26 Cross section, output, and transfer characteristics of four types of MOSFETs.

The device cross sections, output characteristics (i.e., I_D versus V_D), and transfer characteristics (i.e., I_D versus V_G) of the four types are shown in Fig. 26. Note that for the normally off n -channel device, a positive gate bias larger than the threshold voltage V_T must be applied before a substantial drain current flows. For the normally on n -channel device, a large current can flow at $V_G = 0$, and the current can be increased or decreased by varying the gate voltage. This discussion can be readily extended to p -channel device by changing polarities.

5.5.3 Threshold Voltage Control

One of the most important parameters of the MOSFET is the threshold voltage. The ideal threshold voltage is given in Eq. 37. However, when we incorporate the effects of the fixed-oxide charge and the difference in work function, there is a flat-band voltage shift. Additionally, substrate bias can also influence the threshold voltage. When a reverse bias is applied between the substrate and the source, the depletion region is widened and the threshold voltage required to achieve inversion must be increased to accommodate the larger Q_{sc} . These factors in turn cause a change in the threshold voltage:

$$V_T \approx V_{FB} + 2\psi_B + \frac{\sqrt{2\epsilon_s q N_A (2\psi_B + V_{BS})}}{C_o}, \quad (47)$$

where V_{BS} is the reverse substrate-source bias.

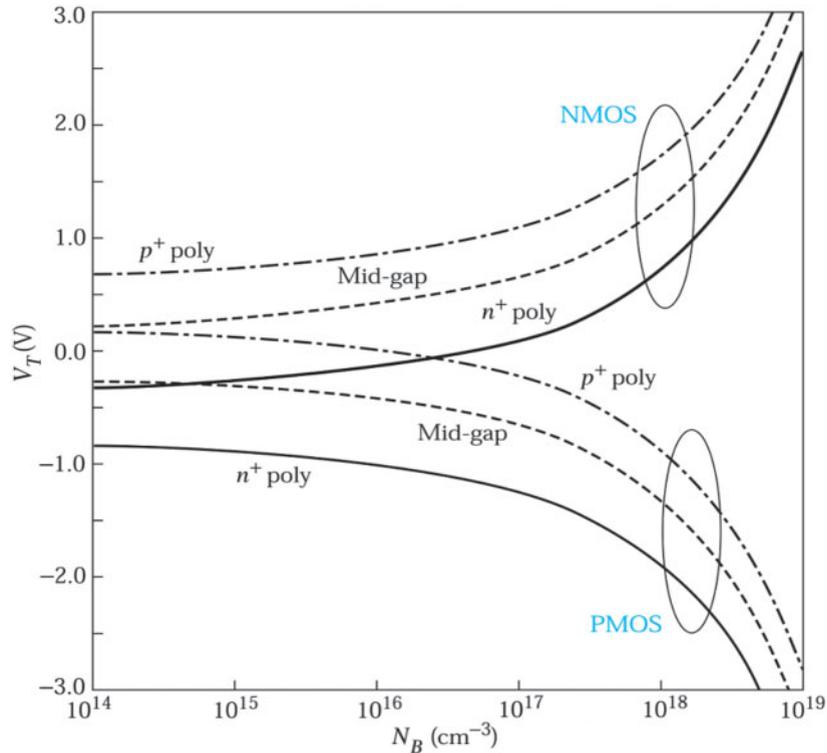


Fig. 27 Calculated threshold voltage of n -channel (V_{Tn}) and p -channel (V_{Tp}) MOSFETs as a function of impurity concentration, for devices with n^+ -, p^+ -polysilicon, and mid-gap work function gates assuming zero fixed charge. The thickness of the gate oxide is 5 nm. NMOS is an n -channel MOSFET; PMOS is a p -channel MOSFET.

Figure 27 shows the calculated threshold voltage of n -channel (V_{Tn}) and p -channel (V_{Tp}) MOSFETs with n^+ -, p^+ -polysilicon and mid-gap work function gate electrodes as a function of their substrate doping, assuming $d = 5$ nm, $V_{BS} = 0$, and $Q_f = 0$. Mid-gap gate materials are those with a work function of 4.61 eV, which equals the sum of the electron affinity $q\chi$ and $E_g/2$ of silicon (see Fig. 2).

Precise control of the threshold voltage of MOSFETs in an integrated circuit is essential for reliable circuit operation. Typically, the threshold voltage is adjusted through ion implantation into the channel region. For example, a boron implantation through a surface oxide is often used to adjust the threshold voltage of an n -channel MOSFET (with p -type substrate). Using this method, it is possible to obtain close control of threshold voltage because very precise quantities of impurity can be introduced. The negatively charged boron acceptors increase the doping level of the channel. As a result, V_T increases. Similarly, a shallow boron implant into a p -channel MOSFET can reduce V_T .

► EXAMPLE 6

For an n -channel n^+ -polysilicon-SiO₂-Si MOSFET with $N_A = 10^{17}$ cm⁻³ and $Q_f/q = 5 \times 10^{11}$ cm⁻², calculate V_T for a gate oxide of 5 nm. What is the boron ion dose required to increase V_T to 0.6 V? Assume that the implanted acceptors form a sheet of negative charge at the Si-SiO₂ interface.

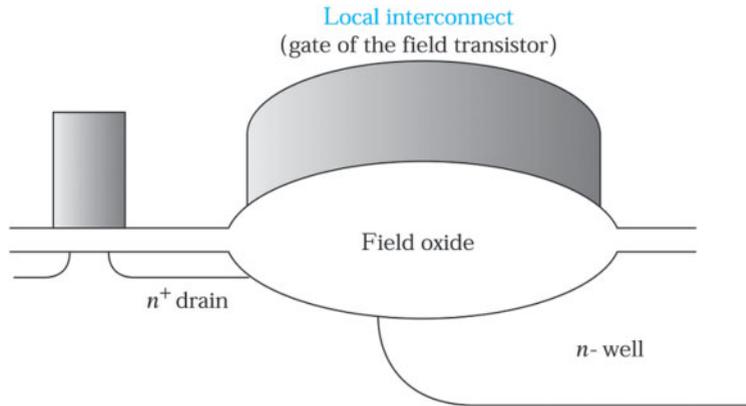


Fig. 28 Cross section of a parasitic field transistor in an n -well structure.

SOLUTION From the examples in Section 5.1, we have $C_o = 6.9 \times 10^{-7} \text{ F/cm}^2$, $2\psi_B = 0.84 \text{ V}$, and $V_{FB} = -1.1 \text{ V}$. Therefore, from Eq. 47 (with $V_{BS} = 0$),

$$\begin{aligned} V_T &= V_{FB} + 2\psi_B + \frac{\sqrt{2\varepsilon_s q N_A (2\psi_B)}}{C_o} \\ &= -1.1 + 0.84 + \frac{\sqrt{2 \times 11.9 \times 8.85 \times 10^{-14} \times 1.6 \times 10^{-19} \times 10^{17} \times 0.84}}{6.9 \times 10^{-7}} \\ &= -0.02 \text{ V.} \end{aligned}$$

The boron charge causes a flat-band shift of qF_B/C_o . Thus,

$$\begin{aligned} 0.6 &= -0.02 + \frac{qF_B}{6.9 \times 10^{-7}}, \\ F_B &= \frac{0.62 \times 6.9 \times 10^{-7}}{1.6 \times 10^{-19}} = 2.67 \times 10^{12} \text{ cm}^{-2}. \end{aligned}$$

We can also control V_T by varying the oxide thickness. Threshold voltage becomes more positive for an n -channel MOSFET and more negative for a p -channel MOSFET as the oxide thickness is increased. This is simply due to the reduced field strength at a fixed gate voltage for a thicker oxide. Such an approach is used extensively for isolating transistors fabricated on a chip. Figure 28 shows the cross section of an isolation oxide (also called field oxide) between an n^+ diffusion and an n -well. Details about the field oxide formation and the well technology are given in Chapter 15. The n^+ diffusion region is the source or drain region of a normal n -channel MOSFET. The gate oxide of MOSFET is much thinner than the field oxide. When a conductor line is formed over the field oxide, a parasitic MOSFET, also called a field transistor, results with the n^+ diffusion and n -well regions as the source and drain, respectively. The V_T of the field oxide is typically an order of magnitude larger than that of the thin gate oxide. During circuit operation, the field transistor will not be turned on. Consequently, the field oxide provides good isolation between the n^+ diffusion and n -well regions.

► EXAMPLE 7

For an n -channel field transistor with $N_A = 10^{17} \text{ cm}^{-3}$ and $Q_f/q = 5 \times 10^{11} \text{ cm}^{-2}$, calculate V_T for a gate oxide (i.e., the field oxide) of 500 nm.

SOLUTION $C_o = \epsilon_{ox} / d = 6.9 \times 10^{-9}$ F/cm².

From Exs. 2 and 3, we have $2\psi_B = 0.84$ V, and

$$V_{FB} = \phi_{ms} - \frac{(Q_f + Q_m + Q_{ot})}{C_o} = -0.98 - \frac{(1.6 \times 10^{-19} \times 5 \times 10^{11})}{6.9 \times 10^{-9}} = -12.98 \text{ V.}$$

Therefore, from Eq. 47 (with $V_{BS} = 0$)

$$\begin{aligned} V_T &= V_{FB} + 2\psi_B + \frac{\sqrt{2\epsilon_s q N_A (2\psi_B)}}{C_o} \\ &= -12.98 + 0.84 + \frac{\sqrt{2 \times 11.9 \times 8.85 \times 10^{-14} \times 1.6 \times 10^{-19} \times 10^{17} \times 0.84}}{6.9 \times 10^{-9}} \\ &= 12.24 \text{ V.} \end{aligned}$$

Substrate bias can also be used to adjust the threshold voltage. The source and substrate may not be at the same potential. The p - n junction between source and substrate must be zero or reverse biased. If the V_{BS} is zero, the gate voltage is at the threshold voltage as in Eq. 47, and the surface potential of substrate is $2\psi_B$. When a reverse substrate-source bias is applied ($V_{BS} > 0$), the potential of electrons in the channel is raised to be higher than that of the source. The electrons in the channel will be pushed laterally to the source. If the electron density in the channel under heavy inversion condition is kept the same, the gate voltage must be raised to $2\psi_B + V_{BS}$. According to Eq. 47, the change in threshold voltage due to the substrate bias is

$$\Delta V_T = \frac{\sqrt{2\epsilon_s q N_A}}{C_o} (\sqrt{2\psi_B + V_{BS}} - \sqrt{2\psi_B}). \quad (48)$$

If we plot the drain current versus V_G , the intercept at the V_G -axis corresponds to the threshold voltage, Eq. 37. Such a plot is shown in Fig. 29 for three different substrate biases. As the magnitude of the substrate V_{BS} increases from 0 V to 2 V, the threshold voltage also increases from 0.56 V to 1.03 V. The substrate effect can be used to raise the threshold voltage of a marginal enhancement device ($V_T \sim 0$) to a larger value.

► EXAMPLE 8

For the MOSFET discussed in Ex. 6 with V_T of -0.02 V, if the reverse substrate bias is increased from zero to 2 V, calculate the change in threshold voltage.

SOLUTION From Eq. 48,

$$\begin{aligned} \Delta V_T &= \frac{\sqrt{2\epsilon_s q N_A}}{C_o} (\sqrt{2\psi_B + V_{BS}} - \sqrt{2\psi_B}) \\ &= \frac{\sqrt{2 \times 11.9 \times 8.85 \times 10^{-14} \times 1.6 \times 10^{-19} \times 10^{17}}}{6.9 \times 10^{-9}} (\sqrt{0.84 + 2} - \sqrt{0.84}) \\ &= 0.27 \times (1.69 - 0.92) = 0.21 \text{ V.} \end{aligned}$$

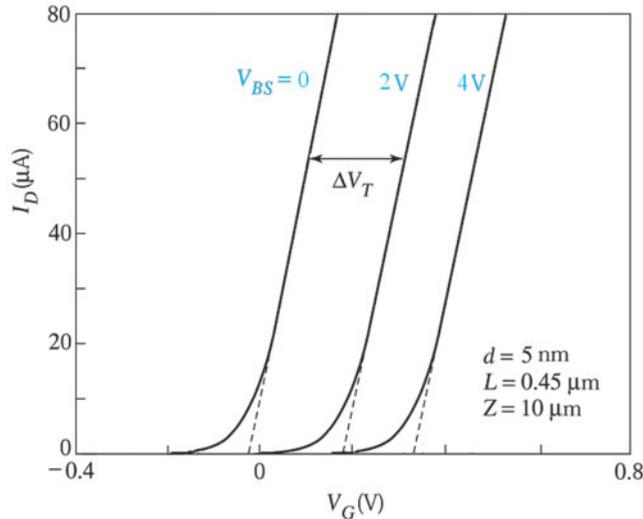


Fig. 29 Threshold voltage adjustment using substrate bias.

Another way to control V_T is to adjust the work function difference by choosing an appropriate gate material. A number of conducting materials have been proposed, such as W, TiN, and a heavily doped polycrystalline silicon-germanium layer.⁹ In deep submicron device fabrication, control of the threshold voltage and device performance becomes more difficult because of the geometric effects encountered in device scaling (see the discussion in the next chapter). The use of other gate materials to replace the conventional n^+ polysilicon could make device design more flexible.

► SUMMARY

In this chapter, we first consider the MOS capacitor, a core component of MOSFET. Charge distributions at the oxide/semiconductor interface (accumulation, depletion, and inversion) in an MOS device can be controlled by the gate voltage. The quality of an MOS capacitor is determined by the qualities of the oxide bulk and oxide/semiconductor interface. For commonly used metal electrodes, the work function difference $q\phi_{ms}$ is generally not zero, and there are various charges inside the oxide or at the SiO_2 -Si interface that will, in one way or another, affect the ideal MOS characteristics. The qualities of the oxide bulk and oxide/semiconductor interface can be evaluated by capacitance-voltage and current-voltage relationships. We then introduced the basic characteristics and the operational principles of the MOSFET. The MOSFET is formed when a source and a drain are placed adjacent to the MOS capacitor. Output current (i.e., drain current) is controlled by varying the gate and drain voltages. The threshold voltage is the main parameter that determines the on-off characteristics of an MOSFET. The threshold voltage can be adjusted by choosing suitable substrate doping, oxide thickness, substrate bias, and gate materials.

► REFERENCES

1. E. H. Nicollian and J. R. Brews, *MOS Physics and Technology*, Wiley, New York, 1982.
2. A. S. Grove, *Physics and Technology of Semiconductor Devices*, Wiley, New York, 1967.
3. B. E. Deal, "Standardized Terminology for Oxide Charge Associated with Thermally Oxidized Silicon," *IEEE Trans, Electron Devices*, ED-27, 606 (1980).

4. S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, 3rd ed., Wiley Interscience, Hoboken, 2007.
5. W. S. Boyle and G. E. Smith, "Charge Couple Semiconductor Devices." *Bell Syst. Tech. J.*, **49**, 587 (1970).
6. M. F. Tompsett, "Video-Signal Generation," in T. P. McLean and P. Schagen, Eds., *Electronic Imaging*, Academic, New York, 1979, p. 55.
7. (a) D. Kahng and M. M. Atalla, "Silicon-Silicon Dioxide Field Induced Surface Devices," *IRE Solid State Device Res. Conf.*, Pittsburgh, PA, 1960. (b) D. Kahng, "A Historical Perspective on the Development of MOS Transistors and Related Devices," *IEEE Trans, Electron Devices*, **ED-23**, 65 (1976).
8. C. C. Hu, *Modern Semiconductor Devices for Integrated Circuits*, Prentice Hall, Upper Saddle River, 2009.
9. Y. V. Ponomarev et al., "Gate-Work Function Engineering Using Poly-(Si,Ge) for High Performance 0.18 μm CMOS Technology," in *Tech. Dig. Int. Electron Devices Meet. (IEDM)*, p.829 (1997).

► **PROBLEMS (* DENOTES DIFFICULT PROBLEMS)**

FOR SECTION 5.1 THE IDEAL MOS CAPACITOR

1. Plot the band diagram of an ideal MOS capacitor with n -type substrate at $V_G = V_T$. Plot the band diagram of an n^+ -polysilicon-gated MOS capacitor with p -type substrate at $V_G = 0$.
2. Plot the band diagram of an n^+ -polysilicon-gated MOS capacitor with p -type substrate at flat-band condition.
3. Plot (a) the charge distribution, (b) electric-field distribution, and (c) potential distribution of an ideal MOS capacitor with n -type substrate under inversion.
5. For a metal-SiO₂-Si capacitor having $N_A = 5 \times 10^{16} \text{ cm}^{-3}$, calculate the maximum width of the surface depletion region.
6. For a metal-SiO₂-Si capacitor having $N_A = 5 \times 10^{16} \text{ cm}^{-3}$ and $d = 8 \text{ nm}$, calculate the minimum capacitance on the C - V curve.
- *7. For an ideal Si-SiO₂ MOS capacitor with $d = 5 \text{ nm}$, $N_A = 10^{17} \text{ cm}^{-3}$, find the applied voltage and the electric field at the interface required to make the silicon surface intrinsic.
8. For an ideal Si-SiO₂ MOS capacitor with $d = 10 \text{ nm}$, $N_A = 5 \times 10^{16} \text{ cm}^{-3}$, find the applied voltage and the electric field at the interface required to bring about strong inversion.

FOR SECTION 5.2 THE SiO₂-Si MOS CAPACITOR

- *9. Assume that the oxide trapped charge Q_{ot} in an oxide layer has a uniform volume charge density, $\rho_{ot}(y)$, of $q \times 10^{17} \text{ cm}^{-3}$, where y is the distance from the location of the charge to the metal-oxide interface. The thickness of the oxide layer is 10 nm. Find the change in the flat-band voltage due to Q_{ot} .
10. Assume that the oxide trapped charge Q_{ot} in an oxide layer is a charge sheet with an area density of $5 \times 10^{11} \text{ cm}^{-2}$ located solely at $y = 5 \text{ nm}$. The thickness of the oxide layer is 10 nm. Find the change in the flat-band voltage due to Q_{ot} .

11. Assume that the oxide trapped charge Q_{ot} in an oxide layer has a triangular distribution: $\rho_{ot}(y) = q \times (5 \times 10^{23} \times y) \text{ cm}^{-3}$. The thickness of oxide layer is 10 nm. Find the change in the flat-band voltage due to Q_{ot} .
12. Assume that initially there is a sheet of mobile ions at the metal-SiO₂ interface. After a long period of electrical stressing under a high positive gate voltage and raised temperature condition, the mobile ions all drift to the SiO₂-Si interface. This leads to a change of 0.3 V in the flat-band voltage. The thickness of oxide layer is 10 nm. Find the area density of Q_m .

FOR SECTION 5.5 MOSFET FUNDAMENTALS

13. Derive Eq. 36 from Eq. 35 in the text assuming $V_D \ll (V_G - V_T)$.
- *14. Derive the I - V characteristics of an MOSFET with the drain and gate connected together and the source and substrate grounded. Can one obtain the threshold voltage from these characteristics?
15. Consider a long-channel MOSFET with $L = 1 \mu\text{m}$, $Z = 10 \mu\text{m}$, $N_A = 5 \times 10^{16} \text{ cm}^{-3}$, $\mu_n = 800 \text{ cm}^2/\text{V}\cdot\text{s}$, $C_o = 3.45 \times 10^{-7} \text{ F/cm}^2$, and $V_T = 0.7 \text{ V}$. Find V_{Dsat} and I_{Dsat} for $V_G = 5 \text{ V}$.
16. Consider a submicron MOSFET with $L = 0.25 \mu\text{m}$, $Z = 5 \mu\text{m}$, $N_A = 10^{17} \text{ cm}^{-3}$, $\mu_n = 500 \text{ cm}^2/\text{V}\cdot\text{s}$, $C_o = 3.45 \times 10^{-7} \text{ F/cm}^2$, and $V_T = 0.5 \text{ V}$. Find the channel conductance for $V_G = 1 \text{ V}$ and $V_D = 0.1 \text{ V}$.
17. For the device described in Prob. 16, find the transconductance.
18. An n -channel, n^+ -polysilicon-SiO₂-Si MOSFET has $N_A = 10^{17} \text{ cm}^{-3}$, $Q_f/q = 5 \times 10^{10} \text{ cm}^{-2}$, and $d = 10 \text{ nm}$. Calculate the threshold voltage.
19. For the device described in Prob. 18, boron ions are implanted to increase the threshold voltage to +0.7 V. Find the implant dose, assuming that the implanted ions form a sheet of negative charges at the Si-SiO₂ interface.
20. A p -channel, n^+ -polysilicon-SiO₂-Si MOSFET has $N_D = 10^{17} \text{ cm}^{-3}$, $Q_f/q = 5 \times 10^{10} \text{ cm}^{-2}$, and $d = 10 \text{ nm}$. Calculate the threshold voltage.
21. For the device described in Prob. 20, boron ions are implanted to decrease the value of threshold voltage to -0.7 V . Find the implant dose, assuming that the implanted ions form a sheet of negative charges at the Si-SiO₂ interface.
22. For the device described in Prob. 20, if the n^+ poly-Si gate is replaced by p^+ poly-Si gate, what will the threshold voltage be?
23. A field transistor with a structure similar to Fig. 28 has $N_A = 10^{17} \text{ cm}^{-3}$, $Q_f/q = 10^{11} \text{ cm}^{-2}$, and an n^+ polysilicon local interconnect as the gate electrode. If the requirement for sufficient isolation between device and well is $V_T > 20 \text{ V}$, calculate the minimum field oxide thickness.
24. An MOSFET has a threshold voltage of $V_T = 0.5 \text{ V}$, a subthreshold swing of 100 mV/decade, and a drain current of 0.1 μA at V_T . What is the subthreshold leakage current at $V_G = 0$?
25. For the device stated in Prob. 24, calculate the reverse substrate-source voltage required to reduce the leakage current by one order of magnitude. ($N_A = 5 \times 10^{17} \text{ cm}^{-3}$, $d = 5 \text{ nm}$).

Advanced MOSFET and Related Devices

- ▶ 6.1 MOSFET SCALING
 - ▶ 6.2 CMOS AND BICMOS
 - ▶ 6.3 MOSFET ON INSULATOR
 - ▶ 6.4 MOS MEMORY STRUCTURES
 - ▶ 6.5 POWER MOSFET
 - ▶ SUMMARY
-

MOSFET is the most important device in modern high-density advanced integrated circuits (IC). We have considered the basic characteristics of the so-called long-channel MOSFET in the previous chapter. Since 1970 the gate length of MOSFETs in production ICs has been scaled down at a rate about 13% per year, and it will continue to shrink in the foreseeable future. The reduction of device dimensions is driven by requirements for both performance and density. In this chapter we will consider some advanced topics on MOSFET scaling, novel scaled device structures, and logic and memory devices.

Specifically, we cover the following topics:

- MOSFET scaling and its associated short-channel effects.
- Complementary MOS (CMOS) logic circuits.
- Silicon-on-insulator devices.
- MOS memory devices.
- Power MOSFETs.

▶ 6.1 MOSFET SCALING

Scaling down of MOSFET's dimensions is a continuous trend since its inception. Smaller device size makes possible higher device density in an integrated circuit. In addition, a smaller channel length improves the driving current ($I_D \sim 1/L$) and thus the operation performance. As a device's dimensions are reduced, however, influences from the side regions of the channel (i.e., source, drain, and isolation edge) become significant. Device characteristics, therefore, deviate from those derived from gradual-channel approximation for long-channel MOSFETs.

6.1.1 Short-Channel Effects

The threshold voltage given in Eq. 47 in Chapter 5 is derived based on the gradual-channel approximation stated in Section 5.5.1. That is, the charges contained in the surface depletion region of the substrate are induced solely from the field created by the gate voltage. In other words, the third term on the right-hand side of Eq. 47 in Chapter 5 is independent of the lateral fields from the source and drain. As channel length is reduced, however,

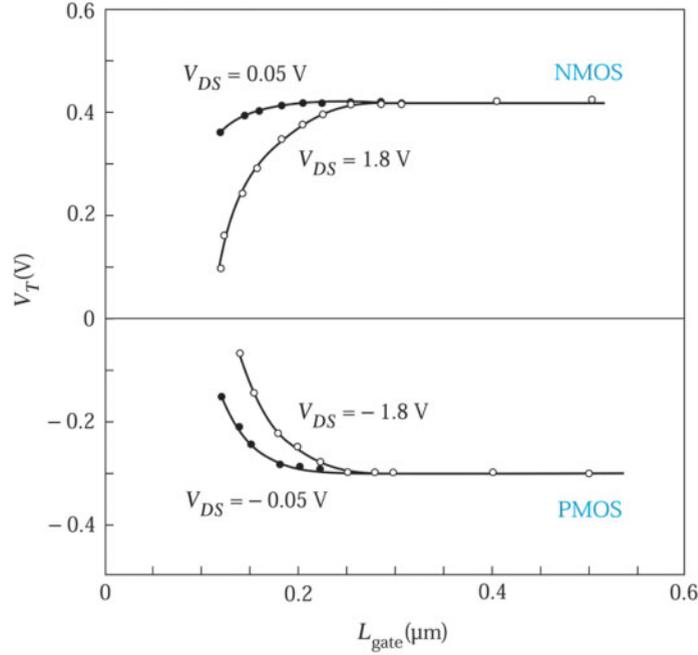


Fig. 1 Threshold voltage roll-off characteristics in a 0.15 μm complementary metal-oxide-semiconductor (CMOS) field-effect transistor technology.¹

the fields originating from the source/drain regions may influence the charge distribution and, thus, device characteristics such as the threshold voltage control and device leakage. When the source and drain depletion regions become a substantial fraction of the channel length, short channel effects start to occur.

Threshold Voltage Roll-off in Linear Region

When short-channel effects become non-negligible, the threshold voltage in the linear region usually becomes less positive as channel length decreases for n -channel MOSFETs and less negative as channel length decreases for p -channel MOSFETs. Figure 1 shows an example of this V_T roll-off phenomenon with $|V_{DS}| = 0.05$ and 1.8 V.¹

Roll-off can be explained by the charge-sharing model.² 2-dimensional examination at the ends of the channel reveals that some of the depletion charge is balanced by the source and drain, as shown in Fig. 2a, in which W_{Dm} is the maximum depletion-layer width, W_S and W_D are the vertical depletion-layer widths under the source and drain. y_S and y_D are the horizontal depletion-layer width at the source and drain ends. $W_D > W_S$ and $y_D > y_S$ are for $V_D > 0$. For small drain bias, we can assume that $W_S \cong W_D \cong W_{Dm}$ as shown in Fig. 2b. The channel depletion region overlaps the source and drain depletion regions, charges induced by the field created by the gate bias can be approximated by those within the trapezoidal region as illustrated in Fig. 2c.

The threshold voltage shift ΔV_T is due to the reduction of charge in the depletion layer from the rectangular region $L \times W_m$ to the trapezoidal region $(L + L') W_m / 2$. ΔV_T is given by (see Prob. 2)

$$\Delta V_T = -\frac{qN_A W_m r_j}{C_o L} \left(\sqrt{1 + \frac{2W_m}{r_j}} - 1 \right), \quad (1)$$

where N_A is the substrate doping concentration, W_m the depletion width, r_j the junction depth, L the channel length, and C_o the gate oxide capacitance per unit area.

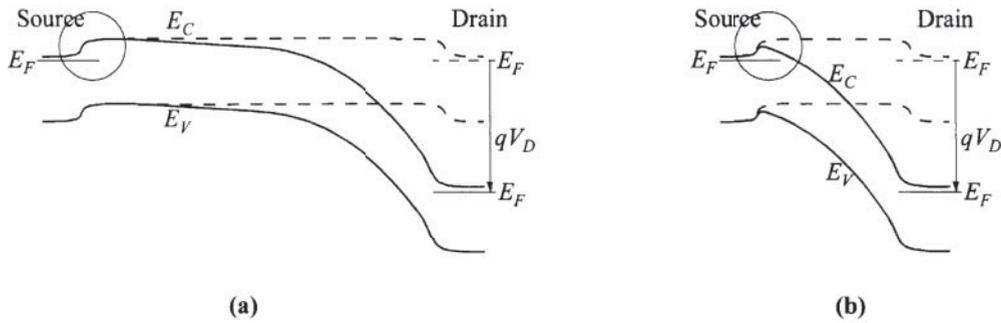


Fig. 3 Energy-band diagram at the semiconductor surface from source to drain, for (a) long-channel and (b) short-channel MOSFETs, showing the DIBL effect in the latter. Dashed lines: $V_D = 0$. Solid lines: $V_D > 0$.

Figure 4 illustrates the subthreshold characteristics of a long and a short n -channel MOSFET at low and high drain bias conditions. The parallel shift in subthreshold current in the short-channel device (Fig. 4b) as the drain voltage increases indicates that a significant DIBL effect has been induced.

Bulk Punch-through

DIBL causes the formation of a leakage path at the SiO_2/Si interface. If the drain voltage is large enough, significant leakage current may also flow from drain to source via the bulk of the substrate for a short-channel MOSFET. This is also ascribed to the increase in the depletion-layer width of the drain junction with increasing drain voltage.

In the extreme case for a short-channel MOSFET, the sum of depletion-layer width for source and drain junctions is comparable to the channel length ($y_s + y_d \cong L$). The depletion region of the drain junction gradually merges with that of the source junction as the drain voltage is increased. An example of severe punch-through characteristics above threshold is shown in Fig. 5a. For this device, at $V_D = 0$ the sum of y_s and y_d is $0.26 \mu\text{m}$, which is larger than the channel length of $0.23 \mu\text{m}$. Therefore, the depletion region of the drain junction has reached the depletion region of the source junction. Over the drain range shown, the device is operated in punch-through condition. Electrons in the source region can be injected into the depleted channel region, where they

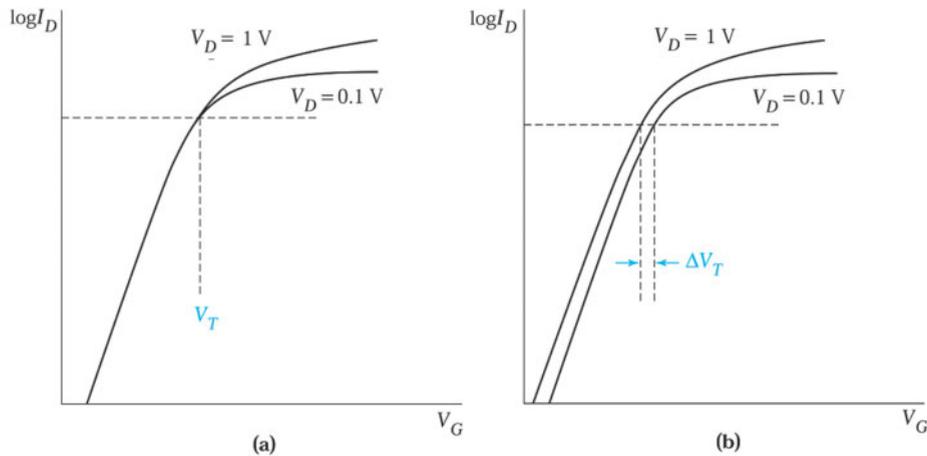


Fig. 4 Subthreshold characteristics of (a) a long-channel and (b) a short-channel MOSFET.

will be swept by the field and collected at the drain, and this leakage current is a strong function of the drain bias. The drain current will be dominated by the space-charge-limited current in the depletion regions:

$$I_D \approx \frac{9\epsilon_s \mu_n A V_D^2}{8L^3}, \quad (2)$$

where A is the cross-sectional area of the punch-through path. The space-charge-limited current increases with V_D^2 and is parallel to the inversion-layer current. The punch-through drain voltage can be estimated by the depletion approximation analogy to Eq. 27, Chapter 3 to be

$$V_{pt} \approx \frac{qN_A(L - y_s)^2}{2\epsilon_s} V_{bi}. \quad (3)$$

The DIBL and bulk punch-through effects on subthreshold current are shown in Fig. 5b for various channel lengths. The device with a 7- μm channel length shows long-channel behavior; that is, the subthreshold drain current is independent of drain voltage. For $L = 3 \mu\text{m}$, there is a substantial dependence of current on V_D , with a corresponding shift of V_T (which is at the point of current departure of the I - V characteristic from the straight line). The subthreshold swing also increases. For an even shorter channel, $L = 1.5 \mu\text{m}$, long-channel behavior is totally lost. The subthreshold swing becomes much worse and the device cannot be turned off.

Figure 6 shows the subthreshold characteristics of a short-channel ($L = 0.23 \mu\text{m}$) MOSFET. When the drain voltage is increased from 0.1 to 1 V, DIBL is induced with the parallel shift in the subthreshold characteristics similar to that shown in Fig. 4b. When the drain voltage is further increased to 4 V, the subthreshold swing is much larger than that for lower drain biases. Consequently, the device has a very high leakage current. This indicates that the bulk punch-through effect is very significant. The gate can no longer turn the device completely off and loses control of the drain current.

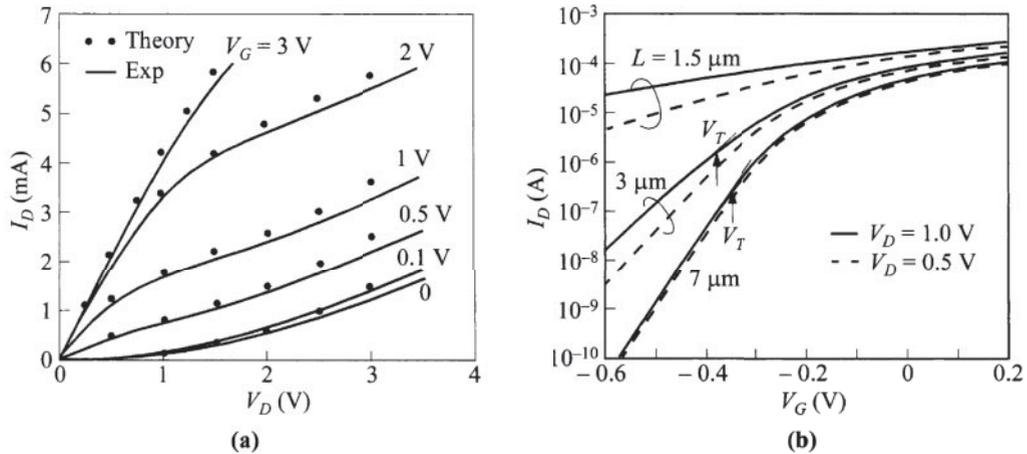


Fig. 5 Drain characteristics of MOSFETs showing punch-through characteristics. (a) Above threshold, $L = 0.23 \mu\text{m}$. $d = 25.8$ nm. $N_A = 7 \times 10^{16}$ cm $^{-3}$. (b) Below threshold. $d = 13$ nm. $N_A = 10^{14}$ cm $^{-3}$.

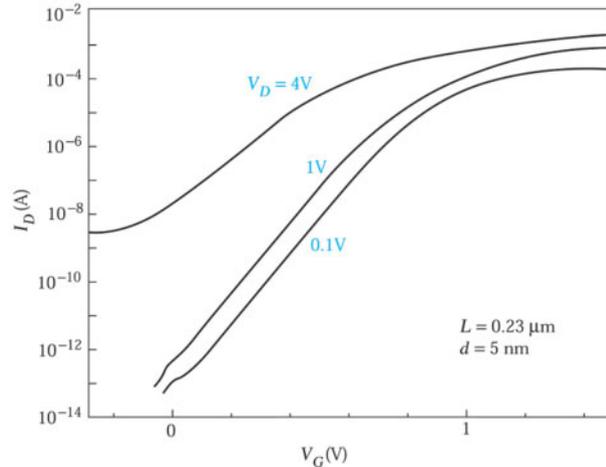


Fig. 6 Subthreshold characteristics of an n -channel MOSFET with $V_D = 0.1, 1,$ and 4 V.

6.1.2 Scaling Rules

As device dimensions are reduced, the short-channel effects must be minimized to maintain normal device and circuit operation. Some guidelines are necessary in scaled-device design. One elegant approach to maintaining the long-channel behavior is to simply reduce all dimensions and voltages by a scaling factor $\kappa (> 1)$, so that the internal electric fields are the same as those of a long-channel MOSFET. This approach is called *constant-field scaling*.³

Table 1 summarizes the scaling rules of the constant-field scaling for various device and circuit parameters.⁴ The circuit performance (speed and power consumption in the on state) can be enhanced as the device dimensions are scaled down.* In practical integrated circuit (IC) manufacturing, however, the electric fields inside smaller devices are not kept constant but increased to some extent. This is mainly because the voltage factors (e.g., power supply, threshold voltage) cannot be scaled arbitrarily. If the threshold voltage is too small, the leakage level in the off-state ($V_G = 0$) will increase significantly because of the nonscalable subthreshold swing. Consequently, standby power consumption will also increase.⁵ By applying the scaling rules, MOSFETs have been fabricated that have a channel length as short as 5 nm, a very low gate delay ($CV/I > 0.22$ ps), high on/off current ratio ($> 5 \times 10^4$), and a reasonable subthreshold swing (~ 75 mV/decade).⁶

TABLE 1 Scaling of MOSFET Device and Circuit Parameters

| Determinant | MOSFET device and circuit parameters | Multiplying factor ($\kappa > 1$) |
|---|--|-------------------------------------|
| Scaling assumptions | Device dimensions (d, L, W, r_f) | $1/\kappa$ |
| | Doping concentration (N_A, N_D) | κ |
| | Voltage (V) | $1/\kappa$ |
| Derived scaling behavior of device parameters | Electric field (\mathcal{E}) | 1 |
| | Carrier velocity (v) | 1 |
| | Depletion-layer width (W) | $1/\kappa$ |
| | Capacitance ($C = \epsilon A/d$) | $1/\kappa$ |
| | Inversion-layer charge density (Q_n) | 1 |

*Figure 19 of Chapter 7 compares the cutoff frequency for different field-effect transistors (including MOSFET).

| | | |
|---|---|--------------|
| Derived scaling behavior of circuit parameters | Current, drift (I) | $1/\kappa$ |
| | Channel resistance (R) | 1 |
| | Circuit delay time ($\tau \sim CVI$) | $1/\kappa$ |
| | Power dissipation per circuit ($P \sim VI$) | $1/\kappa^2$ |
| | Power-delay product per circuit ($P\tau$) | $1/\kappa^3$ |
| | Circuit density ($\sim 1/A$) | κ^2 |
| | Power density (P/A) | 1 |

6.1.3 MOSFET Structures to Control Short-Channel Effects

Many device structures have been proposed to control short-channel effects and improve MOSFET performance⁵. The improvements of an MOSFET structure can be made in three separate parts: channel doping, gate stack, and source/drain design.

Channel Doping Profile

Figure 7 shows the schematic structure of a typical high-performance MOSFET based on planar technology. The channel doping profile has a peak level slightly below the semiconductor surface. This retrograde profile is achieved with ion implantation, often of multiple doses and energies. The low concentration at the surface has the advantages of higher mobility, mainly from the alleviation of surface scattering by normal field reduction due to lower threshold voltage as well as by reduced impurity scattering in the channel. The high peak concentration below the surface is to control punch-through and other short-channel effects. The lower concentration is typically below the junction depth, reducing the junction capacitance as well as the substrate-bias effect on threshold voltage.

Gate Stack

The gate stack consists of the gate dielectric and the gate contact material. As the thickness of SiO_2 for the gate dielectric is scaled into the range below 2 nm, fundamental problems of tunneling and technological difficulty of defects start to demand alternative techniques. High-dielectric materials or high- k dielectrics can have a thicker physical thickness for the same capacitance, thus reducing its electric field. The common terminology used is *the equivalent oxide thickness* ($\text{EOT} = \text{physical thickness} \times k_{\text{SiO}_2}/k$). Some material options being examined are Al_2O_3 , HfO_2 , ZrO_2 , La_2O_3 , Ta_2O_5 , and TiO_2 . The EOT can be easily extended to below 1 nm.

The gate contact material has been polysilicon for a long time. The advantages of a poly-Si are its compatibility with the silicon processing, and its ability to withstand the high-temperature anneal that is

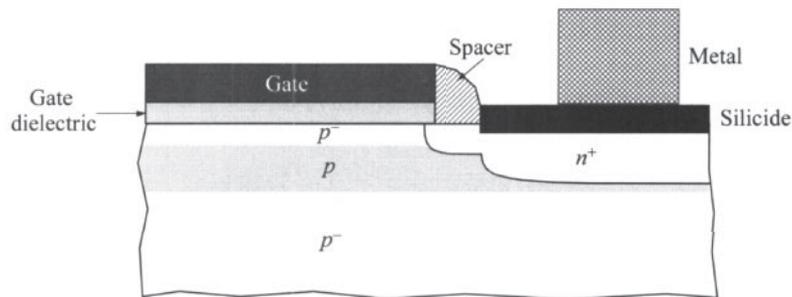


Fig. 7 High-performance MOSFET planar structure with a retrograde channel doping profile, two-step source/drain junction, and self-aligned silicide source/drain contact.

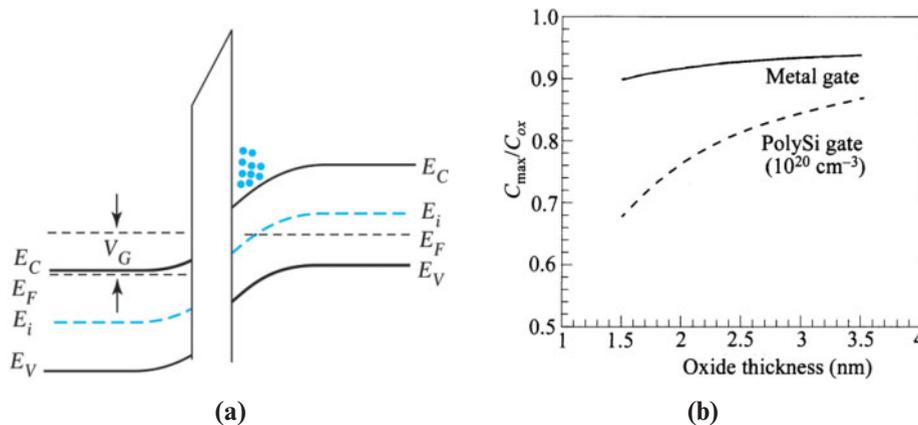


Fig. 8 (a) Band diagram showing poly-Si gate depletion effect of an n^+ -polysilicon-gate n -channel MOS capacitor biased under inversion operation. (b) Degradation in oxide capacitance of MOS capacitor with metal and poly-Si gates.

required after self-aligned source/drain implantation. The self-aligned process can eliminate parasitic capacitances from the overlay errors between gate and source/drain. Another important factor is that the work function can be varied by doping it into n -type and p -type, as shown in Fig. 8, Chapter 5. Such flexibility is crucial for the symmetric CMOS technology. One limitation of the poly-Si gate is its relatively high resistance. It will increase the input impedance and hence lower high-frequency performances. Another shortcoming of poly-Si gate is the depletion effect. Taking an n^+ -polysilicon-gate n -channel MOS capacitor biased under inversion condition for example, as shown in Fig. 8a, the direction of oxide electric field will expel electrons in n^+ polysilicon at the poly-Si/oxide interface. The bands in the n^+ polysilicon bend slightly upward toward the oxide interface to form a depletion region there. The gate depletion results in an additional capacitance in series with the oxide capacitance. This reduces the effective gate capacitance and the inversion-layer charge density to degrade MOSFET transconductance. This becomes more severe with thinner oxides, as shown in Fig. 8b. To circumvent the problems of resistance and depletion, we have to use silicides and metals as the gate contact materials. Potential candidates are TiN, TaN, W, Mo, and NiSi.

Source/Drain Design

As the channel length becomes shorter, the bias voltage must be scaled down accordingly. Otherwise, the increased electrical field could induce avalanche breakdown at the drain. The source/drain structures shown in Fig. 7 have two sections. The extension near the channel has shallower junction depth to minimize short-channel effects. Usually it is doped less heavily (called a *lightly doped drain* (LDD)) to reduce the lateral peak field and to minimize impact ionization by hot carriers in the gate-to-drain overlap region. The deeper junction depth away from the channel helps to minimize the series resistance. The LDD structure can have two disadvantages, the fabrication complexity and the higher drain resistance. However, LDD will result in higher performance.

In the discussion of MOSFET current, the source and drain regions are assumed to be perfectly conducting. Due to the finite silicon resistivity and metal contact resistance, there is a small voltage drop in the source/drain region. In a long-channel MOSFET, the source/drain parasitic resistance is negligible compared with the channel resistance. In a short-channel MOSFET, the source/drain series resistance can be an appreciable fraction of the channel resistance and cause significant current degradation.

A schematic diagram of the current-flow pattern in the source/drain region is shown in Fig. 9. The total source/drain resistance can be divided into several parts: R_{ac} is the accumulation-layer resistance in the gate-source (or -drain) overlap region where the current mainly stays near the surface; R_{sp} is associated with the current spreading from the surface layer into a uniform pattern across the depth of the source/drain; R_{sh} is the sheet resistance of the source/drain region where the current flows uniformly; and R_{co} is the contact resistance in the region where the current flows into a metal line. Once the current flows into an aluminum line, there is very little additional resistance, since the resistivity of aluminum is very low.

There are three ways to reduce the source/drain series resistance.

(a) Silicide Contact Technology

A major milestone for source/drain design is the development of silicide contact technology. A highly conductive silicide film is formed on all the gate and source/drain surfaces separated by dielectric spacers in a *self-aligned process*, as shown in Fig. 7. (This *self-aligned silicide* process has been called *salicide*.) The details of silicide formation will be described in Chapter 12 (Section 12.5.6). Since the sheet resistivity of silicide is 1~2 orders of magnitude lower than that of the source/drain, the silicide layer practically shunts all the currents. Both R_{sh} and R_{co} are greatly reduced. The only significant contribution to R_{sh} is from the nonsilicided region under the spacer.

(b) Schottky-Barrier Source/Drain

Instead of a $p-n$ junction, the use of Schottky-barrier contacts for the source and drain of a MOSFET, as shown in Fig. 10a, can yield some advantages in fabrication and performance. For a Schottky contact, the junction depth can effectively be made zero to minimize the short-channel effects. $n-p-n$ bipolar-transistor action is also absent for undesirable effects such as the bipolar breakdown and latch-up (see Section 6.2.2) phenomena in CMOS circuits. In addition, the elimination of high-temperature implantation can promote better quality in the oxides and better device control.

At thermal equilibrium with $V_G = V_D = 0$, the barrier height of the metal to the p -substrate for holes is $q\phi_{Bp}$ (e.g., 0.84 eV for an ErSi-Si contact), as shown in Fig. 10b. When the gate voltage is above threshold to invert the surface from p -type to n -type, the barrier height between the source and the inversion layer (electrons) is $q\phi_{Bn} = 0.28$ eV, as shown in Fig. 10c. Note that the source contact is reverse-biased under operating conditions (Fig. 10d). For a 0.28 eV barrier, the thermionic-type reverse-saturation current density is of the order of 10^3 A/cm² at room temperature. To increase current density, metals should be chosen to give the highest majority-carrier barrier so the minority-carrier barrier height is minimized, as seen in Eq. 3 in Chapter 7. Additional current due to tunneling through the barrier should help improve the supply of channel carriers. Currently, making the structure on a p -type Si substrate for n -channel MOSFET is more difficult than a p -channel device with n -substrate because metals and silicides that give large barrier heights on p -type silicon are less common.

The disadvantages of the Schottky source/drain are high series resistance due to the finite barrier height, and higher drain leakage current. As also shown in Fig. 10, the metal or silicide contact has to extend underneath the gate for continuity. This process is much more demanding than a junction source/drain which is done by self-aligned implantation diffusion.

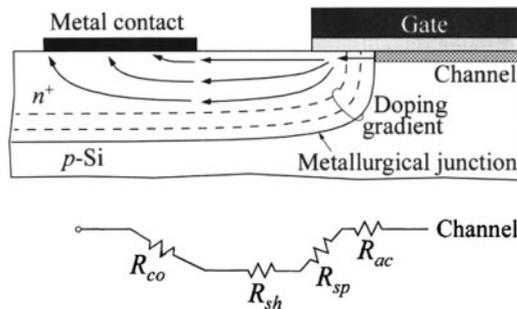


Fig. 9 Detailed analysis of different components of parasitic source/drain series resistance. R_{ac} is the accumulation-layer resistance, R_{sp} is spreading resistance, R_{sh} is the sheet resistance, and R_{co} is the contact resistance.

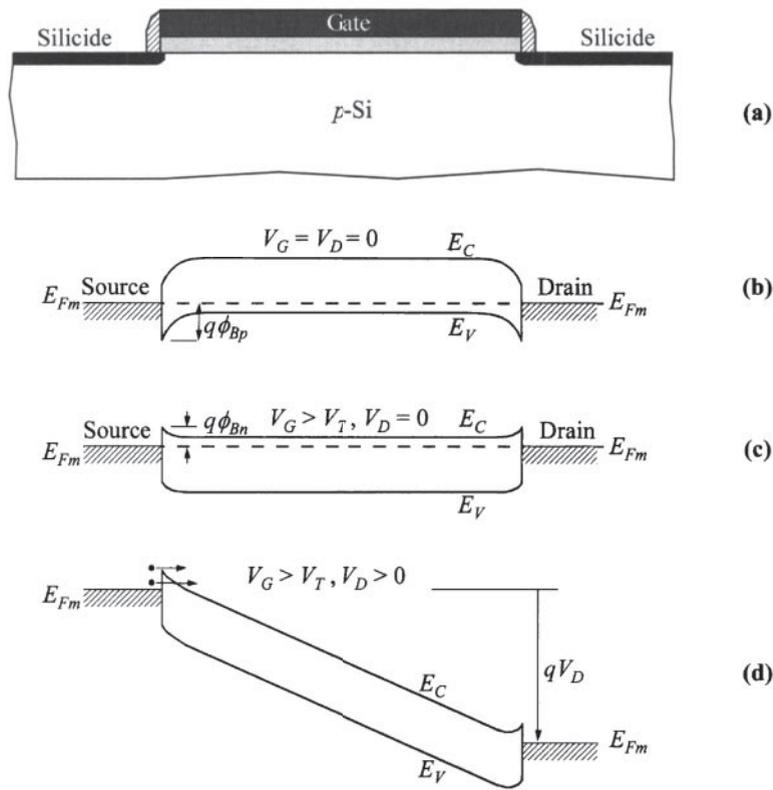


Fig. 10 MOSFET with Schottky-barrier source and drain. (a) Cross-sectional view of the device. (b)-(d) Band diagrams along semiconductor surface under various biases.

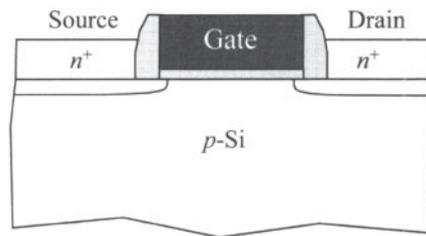


Fig. 11 Raised source/drain to reduce junction depth and series resistance.

(c) Raised Source/Drain

An advanced design is the raised source/drain, in which a heavily doped epitaxial layer is grown over the source/drain regions, as shown in Fig. 11. The purpose is to minimize junction depth to control short-channel effects. Note that an extension underneath the spacer is still needed for continuity.

► 6.2 CMOS AND BiCMOS

Complementary MOS (CMOS) refers to a complementary p -channel and n -channel MOSFET pair. CMOS logic is the most popular technology utilized in present-day integrated circuit design. The main reasons for the success of CMOS are low power consumption and good noise immunity.

6.2.1 The CMOS Inverter

A CMOS inverter, which is the basic element of CMOS logic circuits, is shown in Fig. 12. In a CMOS inverter, the gates of the p - and n -channel transistors are connected and serve as the input node to the inverter. The drains of the two transistors are also connected and serve as the output node to the inverter. The source and substrate contacts of the n -channel MOSFET are grounded, whereas those of the p -channel MOSFET are connected to the power supply (V_{DD}). Note that both p -channel and n -channel MOSFETs are enhancement-type transistors. When the input voltage is low (e.g., $V_{in} = 0$, $V_{GSn} = 0 < V_{Thn}$), the n -channel MOSFET is off.* The p -channel MOSFET, however, is on, since $|V_{GSp}| \cong V_{DD} > |V_{Thp}|$ (V_{GSp} and V_{Thp} are negative). Consequently, the output node is charged to V_{DD} through the p -channel MOSFET. When the input voltage goes high so that the gate voltage equals V_{DD} , the n -channel MOSFET is turned on, since $V_{GSn} = V_{DD} > V_{Thn}$, and the p -channel MOSFET is turned off, since $|V_{GSp}| \cong 0 < |V_{Thp}|$. Therefore, the output node is discharged to ground through the n -channel MOSFET.

For a more detailed understanding of the operation of the CMOS inverter, we can plot the output characteristics of the transistors. This plot is given in Fig. 13, in which I_p and I_n are shown as a function of output voltage (V_{out}). I_p is the current of p -channel MOSFET in the direction from the source (connected to V_{DD}) to the drain (output node). I_n is the current of n -channel MOSFET in the direction from the drain (output node) to the source (connected to ground). Note that the increase in input voltage (V_{in}) tends to increase I_n but decrease I_p at fixed V_{out} . In steady state, however, I_n should be equal to I_p . For a given V_{in} , we can determine the corresponding V_{out} from the intercept of $I_n(V_{in})$ and $I_p(V_{in})$, as shown in Fig. 13. The V_{in} - V_{out} curve, as shown in Fig. 14, is called the transfer curve of the CMOS inverter.⁴

An important characteristic of the CMOS inverter is that when the output is in a steady logic state, i.e., $V_{out} = 0$ or V_{DD} , only one transistor is on. The current flow from the power supply to ground is thus very low and is equal to the leakage current of the off device. In fact, there is significant current conduction only during the short transient period when the two devices are temporarily on. Therefore the power consumption is very low in the static state compared with other types of logic circuits, such as n -channel MOSFETs, bipolar, etc.

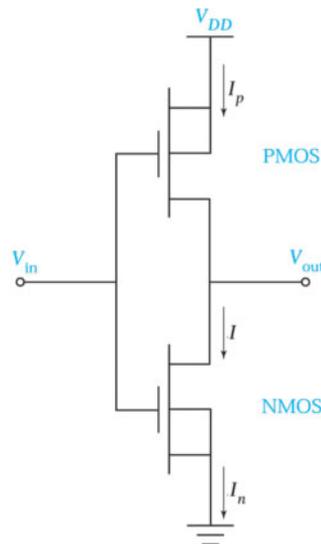


Fig. 12 The CMOS inverter.

* V_{GSn} and V_{GSp} are the voltage differences between the gate and the source for n - and p -channel MOSFETs, respectively.

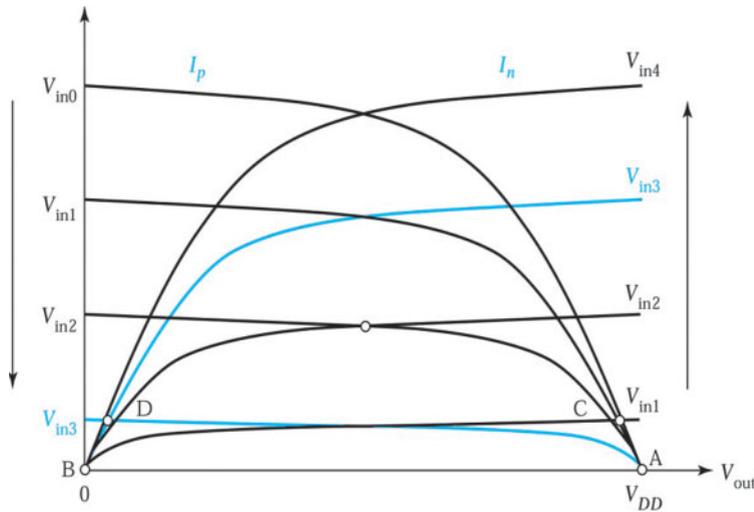


Fig. 13 I_p and I_n as functions of V_{out} . The intercepts of I_p and I_n (circled) represent the steady-state operation points of the CMOS inverter.⁴ The curves are labeled by the input voltages: $0 = V_{in0} < V_{in1} < V_{in2} < V_{in3} < V_{in4} = V_{DD}$.

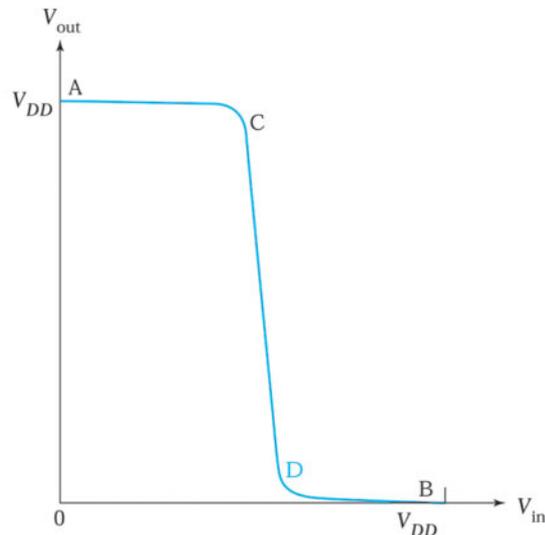


Fig. 14 Transfer curve of a CMOS inverter.⁴ Points labeled A, B, C, and D correspond to the points labeled in Fig. 13.

6.2.2 Latch-up

In order to fabricate both p -channel and n -channel MOSFETs in the same chip for CMOS applications, extra doping and diffusion steps are needed to form the “well” or “tub” in the substrate. The doping type in the well is different from that of the surrounding substrate. Typical well types are the p -well, n -well, and twin well. Details of the well technology are given in Chapter 15. Figure 15 shows a cross-sectional view of a CMOS inverter fabricated using p -well technology. In this figure, the p -channel and n -channel MOSFETs are fabricated in the n -type Si substrate and the p -well region, respectively.

A major problem related to the well structure in CMOS circuits is the latch-up phenomenon. The cause of latch-up is the action of the parasitic p - n - p - n diode in the well structure. As shown in Fig. 15, the parasitic p - n - p - n diode consists of a lateral p - n - p and a vertical n - p - n bipolar transistor. The p -channel MOSFET’s source, n -substrate, and p -well correspond to the emitter, base, and collector of the lateral p - n - p bipolar transistor,

respectively. The n -channel MOSFET's source, p -well, and n -substrate are the emitter, base, and collector of the vertical n - p - n bipolar transistor, respectively. The equivalent circuit of the parasitic components is illustrated in Fig. 16, where R_S and R_W are the series resistance in the substrate and the well, respectively. The base of each transistor is driven by the collector of the other to form a positive feedback loop. This configuration is similar to the thyristor discussed in Chapter 4. Latch-up is induced when the current gain product of the two bipolar transistors, $\alpha_{npn}\alpha_{pnp}$, is larger than 1. When latch-up occurs, a large current will flow from the power supply (V_{DD}) to the ground contact. This can interrupt normal circuit operation and even destroy the chip itself because of the high power dissipation required.

To avoid latch-up, the current gains of the parasitic bipolar transistors must be reduced. One method is to use gold doping or neutron irradiation to lower the minority carrier lifetimes. However, this approach is difficult to control. Besides, it also causes an increase of the leakage current. A deeper well structure or high-energy implantation to form retrograde wells can also reduce the current gain of the vertical bipolar transistor by raising the impurity concentration in the base. In the retrograde well, the peak of the well doping concentration is located within the substrate away from the surface.

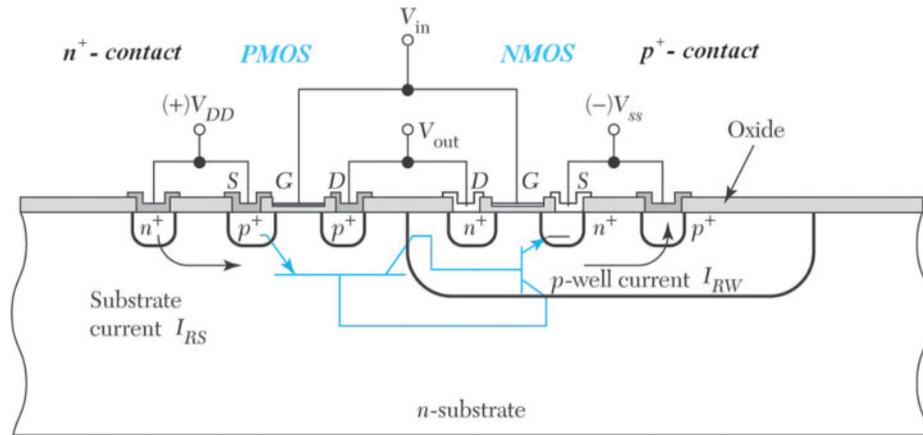


Fig. 15 Cross section of a CMOS inverter fabricated with p -well technology.

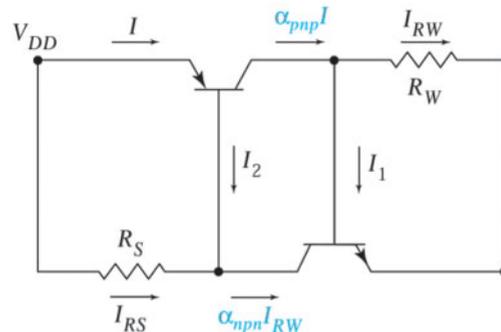


Fig. 16 Equivalent circuit of the p -well structure shown in Fig. 15.

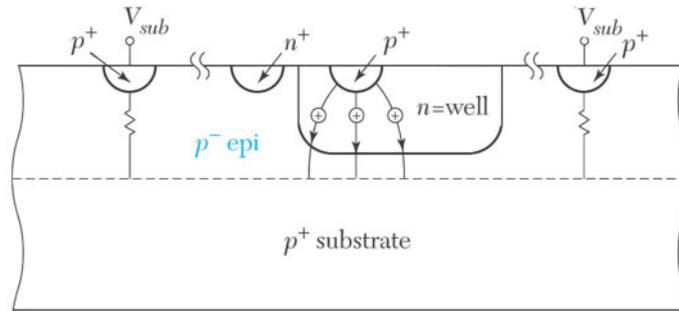


Fig. 17 Prevention of latch-up with a heavily doped substrate.⁷

Another way to reduce latch-up is to use a heavily doped substrate with devices fabricated on a lightly doped epitaxial layer, as shown⁷ in Fig. 17. The heavily doped substrate provides a highly conductive path to collect the current. The current then is drained away through the surface contacts (V_{sub}).

Latch-up can also be avoided with the trench isolation scheme. A process for forming trench isolation is discussed in Chapter 15. This approach can eliminate latch-up because the n -channel and p -channel MOSFETs are physically isolated by the trench.

6.2.3 CMOS Image Sensor

For consumer imaging products such as digital cameras and video recorders, the CCD image sensor discussed in Section 5.4, Chapter 5, dominates the market. However, since the late 1990s⁸ this huge market has been increasingly eroded by the CMOS image sensor fabricated by using standard CMOS processes.

In principle a CMOS image sensor, shown in Fig. 18,⁹ has a very similar architecture to a semiconductor memory. It is composed of an array of identical pixels. Each pixel has a photodiode (a p - n junction photodiode¹⁰), that converts incident light into photocurrent, and an addressing transistor that acts as a switch, as shown in Fig. 19a. A Y-addressing or scan register is used to address the sensor line by line, by activating the in-pixel addressing transistor. An X-addressing or scan register is used to address the pixels on one line, one after another. Some of the readout circuits need to convert the photocurrent into electric charge or voltage and to read it off the array.

The working principle of a pixel is as follows: (1) at the beginning of an exposure the photodiode is reverse biased to a high voltage; (2) during the exposure time, impinging photons decrease the reverse voltage across the photodiode; (3) at the end of the exposure time the remaining voltage across the diode is measured, and the voltage drop from the original value is a measure of the number of photons falling on the photodiode during the exposure time; (4) the photodiode is reset to allow a new exposure cycle.

The most basic form of imaging array shown in Fig. 19b is called PPS (passive pixel sensor), where in each pixel a select transistor controls each photodetector. The advantage is that many cells in a row are accessed at the same time, as in a memory array, so the speed is higher than CCD whose readout is serial in nature. The penalty is larger size.

Many of the differences between CCD and CMOS image sensors arise from differences in their readout architectures. In a CCD (see Fig. 20, Chapter 5), charge is shifted out of the array via vertical and horizontal CCDs, converted into voltage via a simple follower amplifier, and then serially read out. In a CMOS image sensor, charge voltage signals are read out one row at a time in a manner similar to a random-access memory using row and column select circuits.

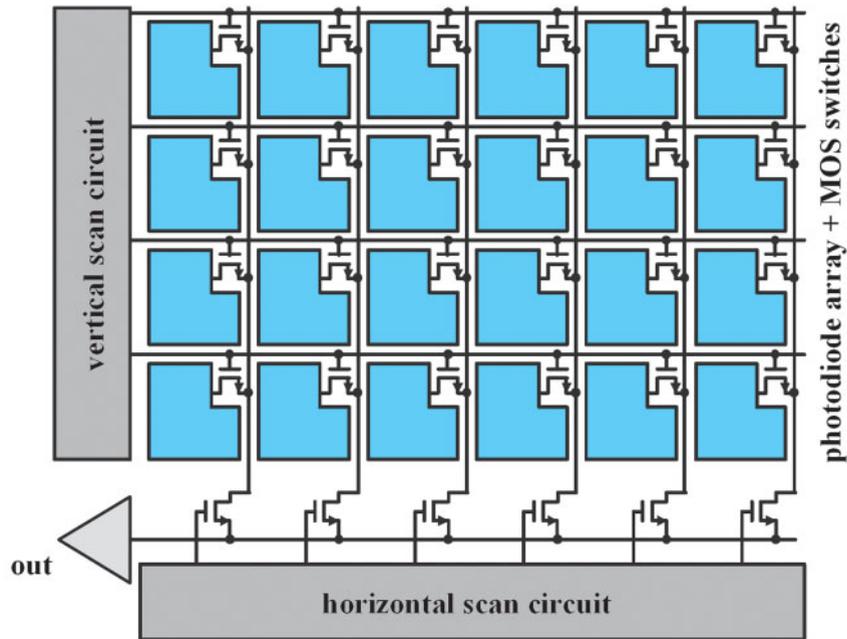


Fig. 18 Architecture of a two-dimensional CMOS image sensor.

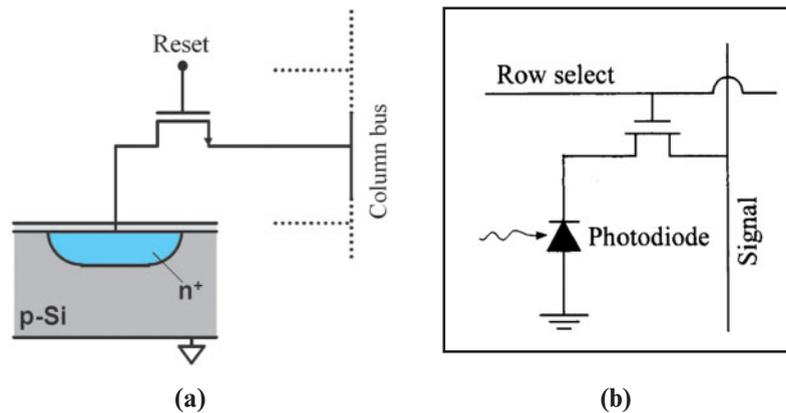


Fig. 19 (a) Passive CMOS pixel based on a single in-pixel transistor. (b) PPS (passive pixel sensor) CMOS image sensors.

The replacement of CCD by CMOS image sensors is growing due to the increasing integration of more functionality within each pixel discussed above, taking advantages of the conventional CMOS scaling and inexpensive technology. Moreover, the advantages of the CMOS image sensor include higher speed due to random-access capability, larger signal-to-noise ratio, lower power due to low voltage requirements, and low cost because of mainstream technology. Conversely, the CCD maintains some advantages such as small pixel size, low-light sensitivity, and high dynamic range. However, CCD requires a different process optimization, so CCD systems that include CMOS circuitry are naturally more expensive.

6.2.4 BiCMOS

CMOS has the advantages of low power dissipation and high device density that make it suitable for fabricating complex circuits. However, CMOS suffers from low drive capability compared with bipolar technology, which limits its circuit performance. BiCMOS is a technology that integrates both CMOS and bipolar device structures in the same chip. A BiCMOS circuit contains mostly CMOS devices, with a relatively small number of bipolar devices. The bipolar devices have better switching performance than their CMOS counterparts without consuming too much extra power. However, this performance enhancement is achieved at the expense of extra manufacturing complexity, longer fabrication time, and higher cost. The fabrication processes for BiCMOS are discussed in Chapter 15.

► 6.3 MOSFET ON INSULATOR

For certain applications, MOSFETs are fabricated on an insulating substrate rather than on a semiconductor substrate. The characteristics of these transistors are similar to those of an MOSFET. Usually, we call such devices thin film transistors (TFT) if the channel layer is an amorphous or polycrystalline silicon. If the channel layer is a monocrystalline silicon, we call it silicon-on-insulator (SOI).

6.3.1 Thin Film Transistor (TFT)

Hydrogenated amorphous silicon (a-Si:H) and polysilicon are the two most popular materials for TFT fabrication. They are usually deposited on an insulating substrate such as a glass, quartz, or Si substrate with a thin SiO₂ capping layer.

The a-Si:H TFT is an important device in electronic applications that require a large area, such as liquid crystal displays (LCD) and contact imaging sensors (CIS). The a-Si:H materials are usually deposited with a plasma-enhanced chemical vapor deposition (PECVD) system. Since the deposition temperature is low (typically 200° – 400°C), inexpensive substrate materials such as glass can be used. The role played by the hydrogen atoms contained in the a-Si:H is to passivate dangling bonds in the amorphous silicon matrix and thus reduce the defect density. Without hydrogen passivation, the gate voltage cannot adjust the Fermi level at the insulator and the a-Si interface, since the Fermi level is pinned by the large amount of defects.

The a-Si:H TFT is usually fabricated using the inverted staggered structure, as shown in Fig. 20. The inverted staggered structure is a bottom-gate scheme. A metal gate can be used since the post-process temperature is low (< 400°C). A dielectric layer such as silicon nitride or silicon dioxide, also deposited by PECVD, is often used as the gate dielectric. An undoped a-Si:H layer is subsequently deposited to form the channel. The source and drain of the TFT are formed with an in situ-doped n⁺ a-Si:H layer complying with the requirement of low process temperature. A dielectric layer that serves as an etch-stop for patterning of n⁺ a-Si:H is often used. Device characteristics of TFTs with the bottom-gate structure are usually better than those with the top-gate structure. This is because the a-Si:H channel could be damaged by plasma during PECVD gate-dielectric deposition of top-gated TFTs. In addition, the source/drain formation process is easier for the bottom-gate structure. A typical subthreshold characteristic of the a-Si:H TFT is shown in Fig. 21. Because of the amorphous matrix present in the channel material, its carrier mobility is usually very low (< 1 cm²/V-s).

The polysilicon TFT uses a thin polysilicon as the channel layer. Polysilicon consists of many Si grains. Within the grains are the monocrystalline Si lattices. The orientations of two side-by-side grains are, however, different from each other. The interface between the two grains is called the grain boundary. Polysilicon TFT exhibits much higher carrier mobility and thus better drive capability than a-Si:H TFT because of higher crystallinity. Carrier mobility of these devices typically ranges from 10 to several hundred cm²/V-s, depending on the grain size and process conditions. Polysilicon is usually deposited with low-pressure CVD (LPCVD). The grain size of polysilicon is an important factor in determining TFT performance, since the carrier mobility generally decreases with decreasing grain size. This is mainly because of the large number of defects contained in the grain boundaries that impede the transport of carriers.

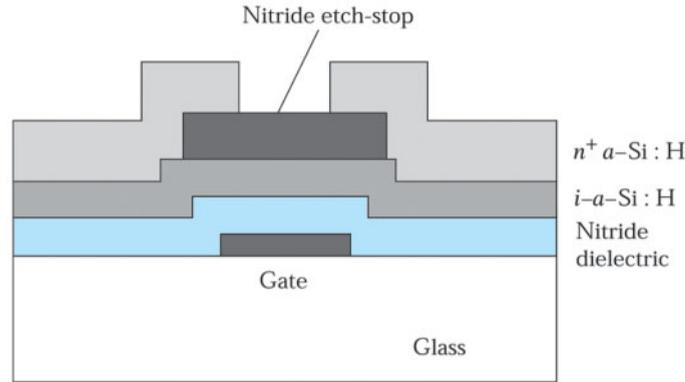


Fig. 20 A typical a-Si:H thin film transistor (TFT) structure.

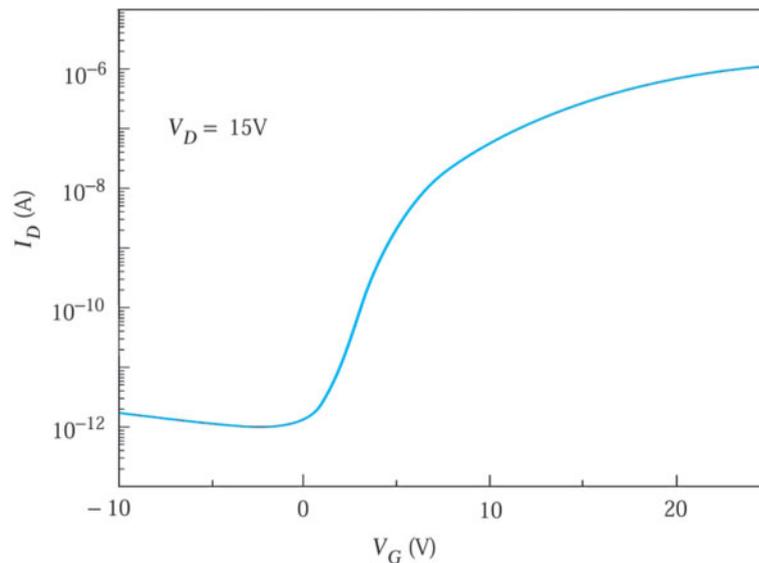


Fig. 21 Subthreshold characteristics of an a-Si:H TFT ($L/Z = 10/60 \mu\text{m}/\mu\text{m}$). The field-effect carrier mobility is $0.23 \text{ cm}^2/\text{V}\cdot\text{s}$.

The defects at the grain boundary can also affect the threshold voltage and subthreshold swing of the device. When gate voltage is applied to induce an inversion layer in the channel, these defects act as traps and impede the movement of the Fermi level in the forbidden gap. To alleviate these drawbacks, a hydrogenation step is often adopted after device fabrication. The hydrogenation treatment is usually done in a plasma reactor. Hydrogen atoms or ions generated in the plasma diffuse into the grain boundaries and passivate these defects. After hydrogenation, there is significant improvement in device performance.

Unlike a-Si:H TFT, polysilicon TFT is usually fabricated with the top-gate structure, as shown in Fig. 22. A self-aligned implant is used to form the source/drain. One main limitation of polysilicon TFT manufacturing is the high process temperature ($> 600^\circ\text{C}$). Consequently, expensive substrates such as quartz are usually needed to tolerate the high process temperatures. This makes polysilicon TFT less attractive than a-Si:H TFT in production for low-end applications because of higher cost. Laser crystallization of Si is a potential way to overcome the problem. In this method, an a-Si layer is deposited first on a glass substrate at low temperatures by PECVD

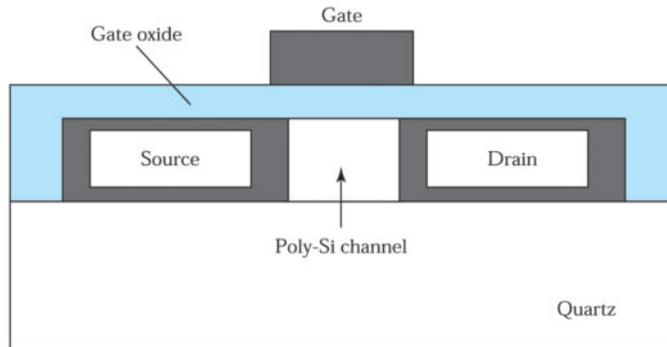


Fig. 22 A polysilicon TFT structure.

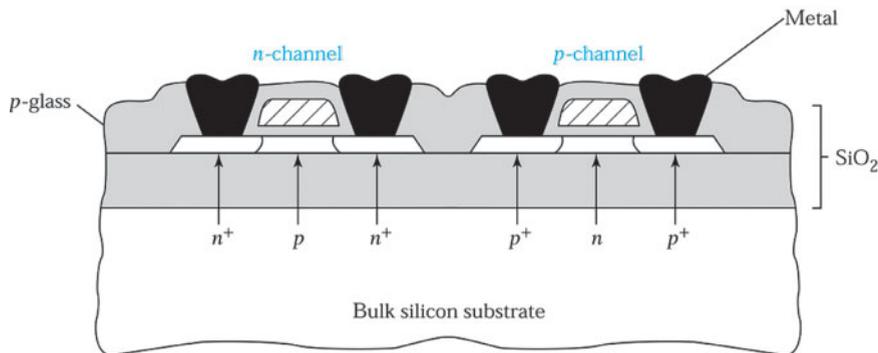


Fig. 23 Cross section of the silicon-on-insulator (SOI).

or LPCVD. A high-power laser source is then used to irradiate the a-Si. The energy is absorbed by the a-Si and melting occurs locally in the a-Si layer. After cooling, the a-Si turns into polysilicon with very large grain size ($\geq 1 \mu\text{m}$). Very high carrier mobility, approaching that of crystalline Si MOSFETs, can be obtained using this method.

6.3.2 Silicon-on-Insulator (SOI) Devices

Many SOI devices have been proposed, including silicon-on-sapphire (SOS), silicon-on-spinel, silicon-on-nitride, and silicon-on-oxide.¹¹ Figure 23 shows a schematic diagram of an SOI CMOS built on silicon dioxide. Compared with CMOS built on a bulk Si substrate (also called bulk CMOS), SOI's isolation scheme is simple and does not need complicated well structures. Device density can thus be increased. The latch-up phenomenon inherent in bulk CMOS circuits is also eliminated. The parasitic junction capacitance in the source and drain regions can be significantly reduced with the insulating substrate. Additionally, significant improvement over bulk CMOS in radiation-damage toleration is achieved in SOI because of the small volume of Si available for electron-hole pair generation by radiation. This property is particularly important for space applications.

Depending on the thickness of the Si channel layer, SOI can be classified into partially depleted (PD) and fully depleted (FD) types. PD-SOI uses a thicker Si channel layer so that the depletion width of the channel does not exceed the thickness of Si layer. Device design and performance of a PD-SOI are similar to that of bulk CMOS. One major difference is the floating substrate used in SOI devices. During device operation, a high field near the drain could induce impact ionization there. Majority carriers (holes in the p -substrate for an n -channel MOSFET) generated by impact ionization will be stored in the substrate, since there is no substrate contact to

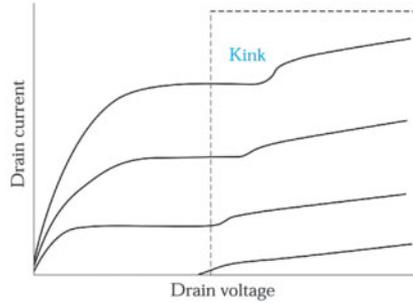


Fig. 24 The kink effect in the output characteristics of an n -channel SOI MOSFET.¹¹

drain away these charges. Therefore, the substrate potential will be changed, which results in a reduction of the threshold voltage. This, in turn, may cause an increase or a kink in the current-voltage characteristics. The kink phenomenon is shown¹¹ in Fig. 24. This float-body or kink effect is especially dramatic for n -channel devices, because of the higher impact-ionization rate of electrons. The kink effect can be eliminated by forming a substrate contact to the source of the transistor. This will, however, complicate the device layout and process flow.

FD-SOI uses a Si layer thin enough so that the channel of the transistor is completely depleted before threshold is reached. This allows the device to be operated at a lower voltage. In addition, the kink effect caused by high-field impact ionization can be eliminated. FD-SOI is very attractive for low-power applications. Nevertheless, the FD-SOI's characteristics are sensitive to variation in the Si thickness. If an FD-SOI circuit is built on a wafer with nonuniform Si thickness, its operation will be unstable.

► EXAMPLE 1

Calculate the threshold voltage for an n -channel SOI device having $N_A = 10^{17} \text{ cm}^{-3}$, $d = 5 \text{ nm}$, and $Q_f/lq = 5 \times 10^{11} \text{ cm}^{-2}$. Si thickness, d_{Si} , for the device is 50 nm.

SOLUTION From Ex. 1, Chapter 5, the maximum depletion width, W_m , for a bulk NMOS device is 100 nm. Therefore, the SOI device is a fully depleted type. Since the width of the depletion region is now the Si thickness, W_m used in Eq. 17 and Eq. 47, Chapter 5 for calculating the threshold voltage should be replaced by d_{Si} :

$$V_T = V_{FB} + 2\psi_B + \frac{qN_A d_{\text{Si}}}{C_o}$$

From Exs. 2 and 3, Chapter 5, we have $C_o = 6.9 \times 10^{-7} \text{ F/cm}^2$, $V_{FB} = -1.1 \text{ V}$, and $2\psi_B = 0.84 \text{ V}$.

Therefore,

$$V_T = -1.1 + 0.84 + \frac{1.6 \times 10^{-19} \times 10^{17} \times 5 \times 10^{-6}}{6.9 \times 10^{-7}} = -0.14 \text{ V.}$$

6.3.3 Three-Dimensional Structures

In device scaling, the optimum design entails MOSFET built on a body of an ultra-thin layer so that the body is fully depleted under the whole bias range. A design to achieve this more efficiently is to have a surround gate structure that encloses the body layer from at least two sides. Two examples of these three-dimensional structures are shown in Fig. 25. They can be classified according to their current-flow pattern: the horizontal transistor¹² (FinFET, the fabrication process, introduced in Chapter 15) and the vertical transistor¹³. Both of these are very challenging from a fabrication point of view. The horizontal transistor is more compatible with SOI technology.

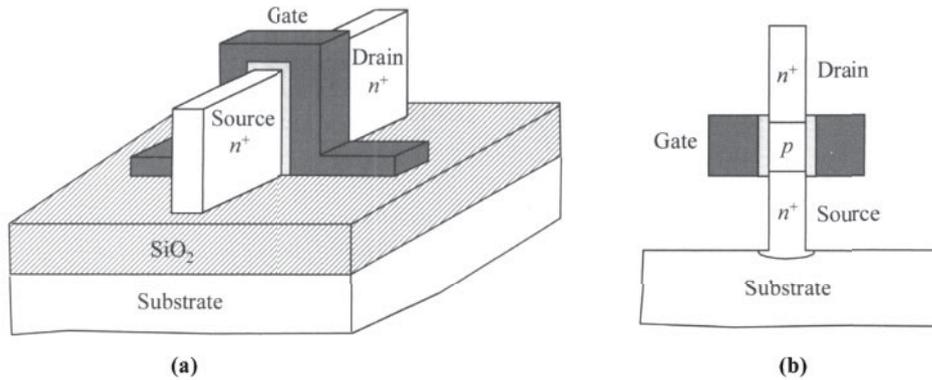


Fig. 25 Schematic three-dimensional MOSFETs. (a) Horizontal structure. (b) Vertical structure.

A set of difficulties arise because the majority or all of the channel surface is on a vertical wall for both of these structures. This presents great challenges in achieving a smooth channel surface from etching and growth or deposition of gate dielectrics on these surfaces. Formation of the source/drain junction is no longer trivial by means of ion implantation. Salicide formation will also be much more difficult. Whether one of these turns out to be the future device of choice remains to be seen.

► 6.4 MOS MEMORY STRUCTURES

Semiconductor memories can be classified as volatile and nonvolatile. Volatile memories such as dynamic random-access memories (DRAMs) and static random-access memories (SRAMs) lose their stored information if the power supply is switched off. Nonvolatile memories, on the other hand, can retain the stored information. Currently, DRAM and SRAM are extensively used in personal computers and workstations, mainly because of DRAM's attributes of high density and low cost and SRAM's attribute of high speed. The nonvolatile memory is used extensively in portable electronics systems such as the cellular phone, digital camera, and smart IC cards, mainly because of its attributes of low-power consumption and nonvolatility.

6.4.1 DRAM

Modern DRAM technology consists of a cell array using the storage cell structure shown¹⁴ in Fig. 26. The cell includes an MOSFET and an MOS capacitor [i.e., one transistor/one capacitor (1T/1C) cell]. The MOSFET acts as a switch to control the writing, refreshing, and read-out actions of the cell. The capacitor is used for charge storage. During the write cycle, the MOSFET is turned on so that the logic state in the bit line is transferred to the storage capacitor. For practical applications, charges stored in the capacitor will be gradually lost because of the small but nonnegligible leakage current of the storage node. Consequently, the operation of DRAM is "dynamic," since the data need to be refreshed periodically within a fixed interval, typically 2–50 ms.

The 1T/1C DRAM cell has the advantages of very simple and small area construction. In order to increase the storage density of a chip, aggressive scaling of the cell size is necessary. However, this will degrade the storage capability of the capacitor, since the capacitor electrode area will be diminished as well. To solve this problem, three-dimensional (3-D) capacitor structures are required. Some novel 3-D capacitor structures are discussed in Chapter 15. High dielectric-constant materials can also be used to replace conventional oxide-nitride composite layers (dielectric constant: 4 ~ 6) as the capacitor dielectric materials in order to increase the capacitance.

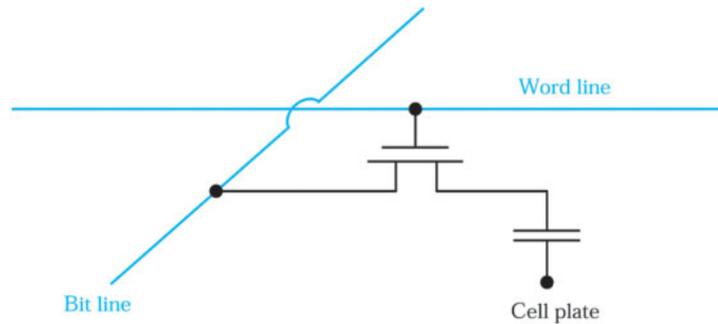


Fig. 26 Basic configuration of a dynamic random-access memory (DRAM) cell.¹⁴

6.4.2 SRAM

SRAM is a matrix of static cells using a bistable flip-flop structure to store the logic state, as shown in Fig. 27. The flip-flop consists of two cross-coupled CMOS inverters (T1, T3 and T2, T4). The output of the inverter is connected to the input node of the other inverter. This configuration is called “latched.” Two additional *n*-channel MOSFETs, T5 and T6, with their gates connected to the word line, are used to access the SRAM cell. The operation of the SRAM is static since the logic state is sustained as long as the power is applied. Therefore, SRAM does not have to be refreshed. The two *p*-channel MOSFETs (T1 and T2) in the inverters are used as the load transistors. There is essentially no dc current flow through the cell, except during switching. In some situations, *p*-channel polysilicon TFTs or polysilicon resistors are used instead of bulk *p*-channel MOSFETs. These polysilicon load devices can be fabricated over the bulk *n*-channel MOSFETs. 3-D integration can effectively reduce the cell area and thus increase the storage capacity of the chip.

6.4.3 Nonvolatile Memory

When the gate electrode of a conventional MOSFET is modified so that semipermanent charge storage inside the gate is possible, the new structure becomes a nonvolatile memory device. Since the first nonvolatile memory device was proposed¹⁵ in 1967, various device structures have been made. Nonvolatile memory devices have been used extensively in ICs such as the erasable-programmable read-only memory (EPROM), electrically erasable-programmable read-only memory (EEPROM), and flash memory.

There are two groups of nonvolatile memory devices, floating-gate devices and charge-trapping devices (Fig. 28). In both types of devices, charges are injected from the silicon substrate across the first insulator and stored in the floating gate or in the nitride. The stored charges give rise to a threshold-voltage shift, and the device is switched to a high-threshold state (programmed or logical 1). In a well-designed memory device, the charge retention time can be over 100 years. To return to the low-threshold state (erased or logical 0), a gate or other means (such as ultraviolet light) can be applied to erase the stored charges.

Floating-Gate Devices

In floating-gate memory devices, charge is injected to the floating gate to change the threshold voltage. The programming can be done by either hot carrier injection or a Fowler-Nordheim tunneling process. Figure 29*a*, which is the same as Fig. 28*b*, shows the hot electron injection scheme in an *n*-channel floating-gate device. Near the drain, the lateral field is at its highest level. The channel electrons acquire energy from the field and become hot electrons. Some of the hot electrons with energy higher than the barrier height of SiO₂/Si conduction band (~3.2 eV) can surmount the barrier and are injected into the floating gate. At the same time, the high field also induces impact ionization. These generated secondary hot electrons can also be injected to the floating gate. Figure 29*b* and *c* show the band diagrams of a floating-gate device under programming and erasing conditions, respectively.

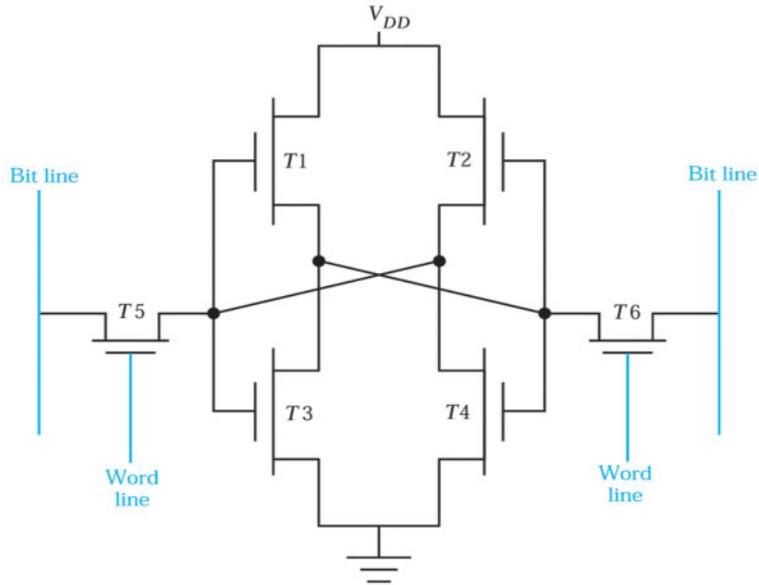


Fig. 27 Configuration of a CMOS SRAM cell. T1 and T2 are load transistors (*p*-channel), T3 and T4 are drive transistors (*n*-channel), and T5 and T6 are access transistors (*n*-channel).

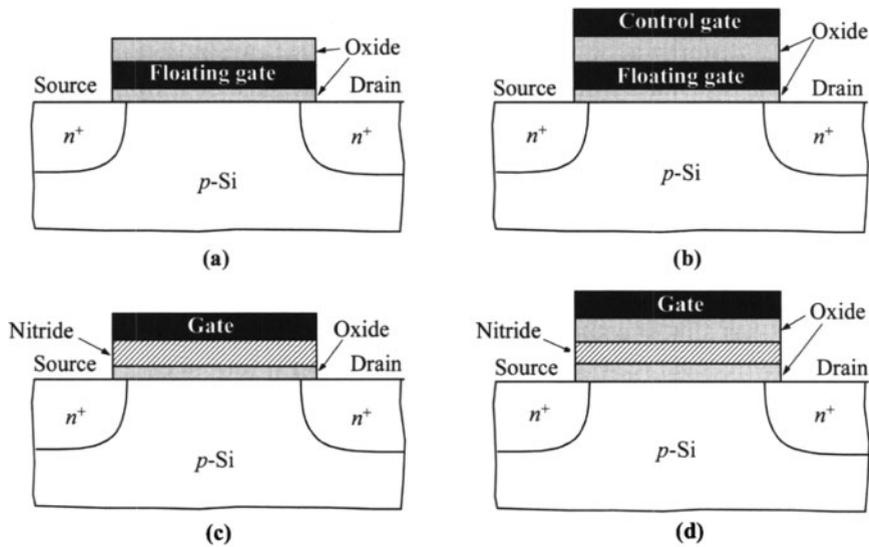


Fig. 28 Variations of nonvolatile memory devices: Floating-gate devices as (a) FAMOS transistor and (b) stacked-gate transistor; charge-trapping devices as (c) MNOS transistor and (d) SONOS transistor.

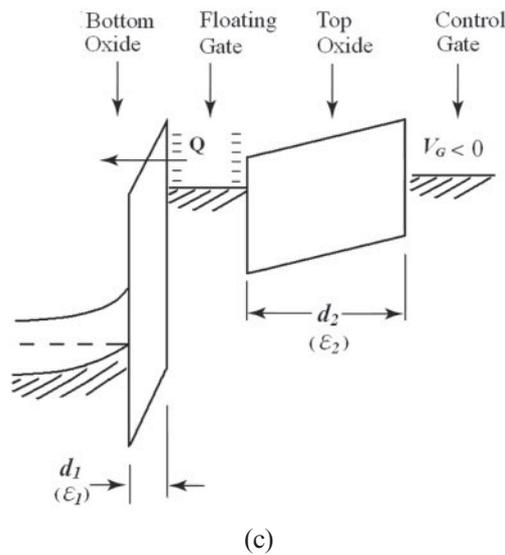
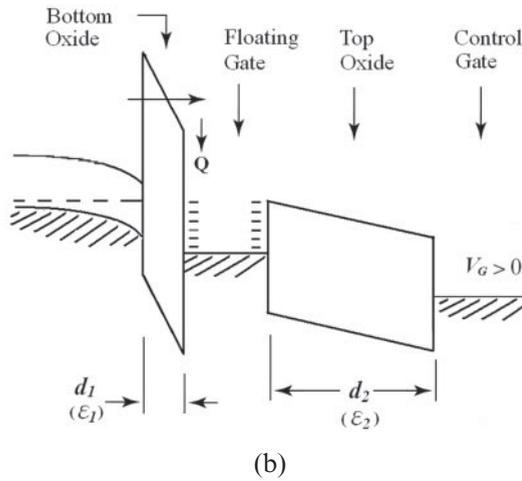
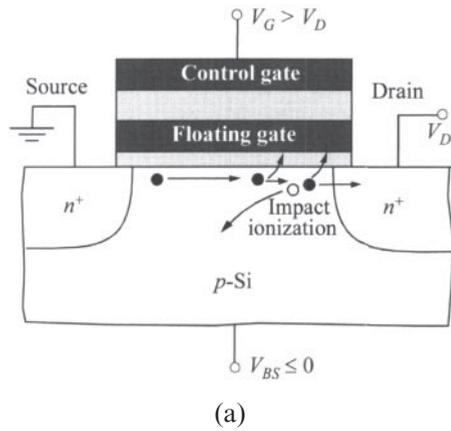


Fig. 29 (a) Charging of the floating gate by hot electrons from channel and impact ionization. The band diagrams of floating-gate device in (b) programming condition and (c) erasing condition.

In the programming mode, the electric field across the bottom oxide layer is most critical. On application of a positive voltage V_G to the control gate, an electric field is established in each of the two dielectrics. We have, from Gauss' law, that (assuming the voltage drop in the semiconductor is small):

$$\varepsilon_1 \mathcal{E}_1 = \varepsilon_2 \mathcal{E}_2 + Q \quad (4)$$

and

$$V_G = V_1 + V_2 = d_1 \mathcal{E}_1 + d_2 \mathcal{E}_2, \quad (5)$$

where the subscripts 1 and 2 correspond to the bottom and top oxide layer respectively, and Q (negative) is the stored charge in the floating gate. In practical devices, the bottom layer has a tunnel oxide of ~ 8 nm, while the top insulator stack typically has an equivalent oxide thickness of ~ 14 nm.

From Eqs. 4-5 we obtain

$$\mathcal{E}_1 = \frac{V_G}{d_1 + d_2(\varepsilon_1 / \varepsilon_2)} + \frac{Q}{\varepsilon_1 + \varepsilon_2(d_1 / d_2)}. \quad (6)$$

The current transport in insulators is generally a strong function of the electric field. When the transport is Fowler-Nordheim tunneling, the current density has the form

$$J = C \varepsilon_1^2 \exp\left(\frac{D}{\mathcal{E}_1}\right), \quad (7)$$

where C and D are constants in terms of effective mass and barrier height.

After charging, the total stored charge Q is equal to the integrated injection current. This causes a shift of the threshold voltage by the amount

$$\Delta V_T = -\frac{d_2 Q}{\varepsilon_2}. \quad (8)$$

The threshold-voltage shift can be measured directly as shown in the I_D - V_G plots (Fig. 30). Alternately, this threshold-voltage shift can be measured from the drain conductance. For small drain voltage, the channel conductance of an n -channel MOSFET is given by

$$g_D = \frac{I_D}{V_D} = \frac{Z}{L} \mu C_{ox} (V_G - V_T). \quad (9)$$

The change in V_T results in a change in the channel conductance g_D . The g_D - V_G plot shifts to the right by ΔV_T .

To erase the stored charge, a negative bias is applied to the control gate or a positive bias to the source/drain. The process is the reverse of the programming process, and the stored electrons tunnel out of the floating gate to the substrate.

Figure 28a shows a floating-gate memory device without the control gate. The first EPROM was developed using heavily doped polysilicon as the floating-gate material. The polysilicon gate is embedded in oxide and is completely isolated. Similar to the stacked-gate transistor shown in Fig. 29, the drain junction is biased to avalanche breakdown, and electrons in the avalanche plasma are injected from the drain region into the floating

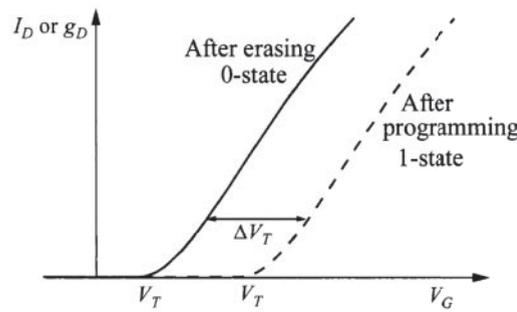


Fig. 30 Drain-current characteristics of a stacked-gate n -channel memory transistor, showing the change of threshold voltage after erasing and programming.

gate. This device is known as a floating-gate avalanche-injection MOS memory (FAMOS). To erase the FAMOS memory, ultraviolet light or x-ray is used, which can excite the stored charges into the conduction band of the gate oxide and back to the substrate. Electrical erasing cannot be used because the device has no external gate.

Flash Memory

Several types of floating-gate devices are differentiated by the erase mechanisms. In the EPROM, which has only a floating gate and no control gate, erasing is done by UV irradiation. The EPROM has the advantage of small cell area due to the 1T (one storage transistor)/cell structure. Nevertheless, its erase scheme necessitates the use of an expensive package with a quartz window. In addition, the erasing time is long.

The EEPROM uses the tunneling process to erase the stored charges. Unlike the EPROM device, in which all cells are erased during erasing, a cell in an EEPROM can be erased only when it is “selected.” This function is accomplished through the selection transistor contained in each cell. Such “bit-erasable” characteristics make the EEPROM more flexible. However, the 2T (one selection transistor plus one storage transistor)/cell feature of EEPROM limits its storage capacity.

The cell structure of flash memory consists of three layers of polysilicon, as shown in Fig. 31.¹⁶ The cell is programmed by a channel hot carrier injection mechanism similar to EPROM. Erasing is accomplished by field emission of electrons from the floating gate to an erase gate. The erase gate is supplied with a boosted voltage that makes possible field emission from the floating gate. The erasing speed is much faster than that of EPROM, whence the name “flash.” The storage cells for a flash memory are divided into several sectors (or blocks). The erasing scheme is performed on one selected sector with the tunneling process. During erasing, all cells in the selected sector are erased simultaneously. The third polysilicon layer is used both as a gate of the selection transistor and a control gate of the cell, and the 1T/cell feature makes the storage capacity of the flash memory higher than that of EEPROM.

Single-electron Memory Cell

A related device structure is the single-electron memory cell (SEMC), which is a limiting case of the floating-gate structure¹⁷. By reducing the length of the floating gate to ultrasmall dimensions, say 10 nm, we obtain the SEMC. A cross-sectional view of an SEMC is shown in Fig. 32. The floating dot corresponds to the floating gate in Fig. 28b. Because of its small size the capacitance is also very small (~ 1 aF). When an electron tunnels into the floating dot, because of the small capacitance, a large tunneling barrier arises to prevent the transfer of another electron. SEMC is the ultimate floating-gate memory cell, since we need only one electron for information storage. A single-electron memory with densities as high as 256 terabits (256×10^{12} bits) that can operate at room temperature has been projected.

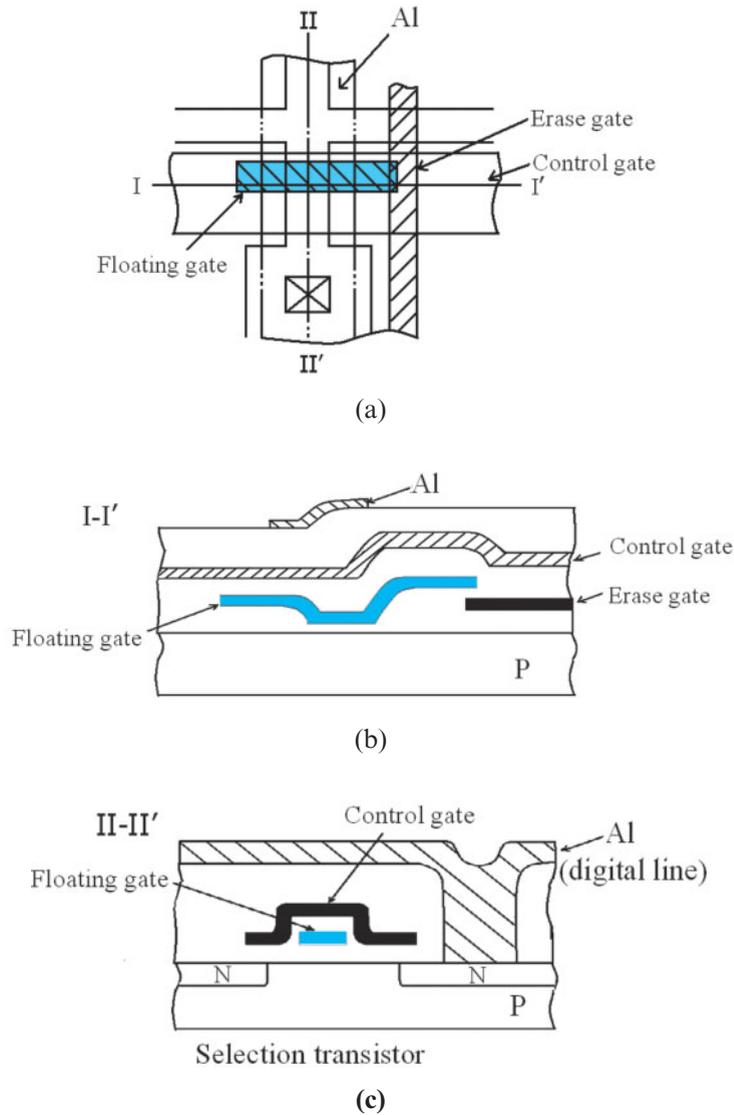


Fig. 31 (a) Top view of flash memory. (b) Cross-sectional view along I-I' line in (a). (c) Cross-sectional view along II-II' line in (a).¹⁶

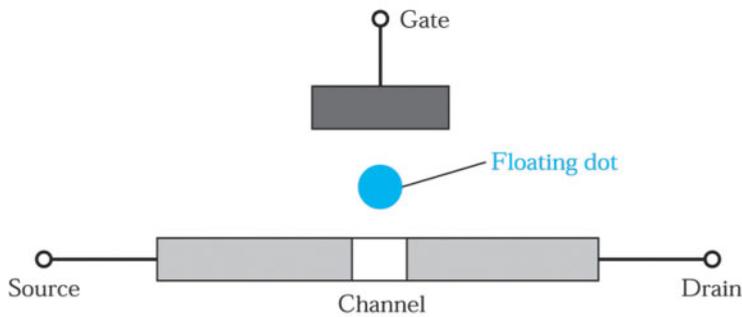


Fig. 32 Illustration of single-electron memory cell.¹⁷

Charge-Trapping Devices

MNOS Transistor

In the MNOS transistor as shown in Fig. 28c, the silicon-nitride layer is used as an efficient material to trap electrons as current passes through the dielectric. Other insulators instead of the silicon-nitride film, such as aluminum oxide, tantalum oxide, and titanium oxide, have been used but are not as common. Electrons are trapped in the nitride layer close to the oxide-nitride interface. The function of the oxide is to provide a good interface to the semiconductor and to prevent back-tunneling of the injected charge for better charge retention. Its thickness has to be balanced between retention time and programming voltage and time.

Figure 33 shows the basic band diagram for the programming and erasing operations. In the programming process, a large positive bias is applied to the gate. Electrons are emitted from the substrate to the gate. The current conduction mechanisms in the two dielectric layers are very different. The current through the oxide is by electrons tunneling through the trapezoidal oxide barrier, followed by a triangular barrier in the nitride. This form of tunneling has been identified as modified Fowler-Nordheim tunneling, as opposed to Fowler-Nordheim tunneling through a single triangular barrier. Then, electrons pass through the nitride layer by Frenkel-Poole transport. When the negative charge starts to build up, the oxide field decreases and the modified Fowler-Nordheim tunneling starts to limit the current.

The threshold voltage is shown in Fig. 34 as a function of programming pulse width. Initially, the threshold voltage changes linearly with time, followed by a logarithmic dependence, and finally it tends to saturate. This programming speed is largely affected by the choice of oxide thickness: a thinner oxide allows shorter programming time. Programming speed must be balanced with charge retention time, since too thin an oxide will allow the trapped charge to tunnel back to the silicon substrate.

The total gate capacitance C_G of the dual dielectrics is equal to the serial combination of their capacitances:

$$C_G = \frac{1}{(1/C_n) + (1/C_{ox})} = \frac{C_{ox}C_n}{C_{ox} + C_n}, \quad (10)$$

where the capacitances $C_{ox} = \epsilon_{ox}/d_{ox}$ and $C_n = \epsilon_n/d_n$ correspond to the oxide and nitride layers, respectively. The amount of trapped charge density Q near the nitride-oxide interface depends on the trapping efficiency of the nitride. The final threshold-voltage shift is given by

$$\Delta V_T = \frac{Q}{C_n}. \quad (11)$$

In the erasing process, a large negative bias is applied to the gate (Fig. 33b). Traditionally, the discharge process was believed to be due to the tunneling of trapped electrons back to the silicon substrate. New evidence shows that the major process is due to tunneling of holes from the substrate to neutralize the trapped electrons. The discharge process as a function of pulse width is also shown in Fig. 34.

The advantages of the MNOS transistor include reasonable speed for programming and erasing, so it is a candidate as a nonvolatile RAM device. It also has superior radiation hardness, due to minimal oxide thickness and the absence of a floating gate. The drawbacks of the MNOS transistor are large programming and erasing voltages and nonuniform threshold voltage from device to device. The passage of tunneling current gradually increases the interface-trap density at the semiconductor surface and also causes a loss of trapping efficiency due to leakage or tunneling of trapped electrons back to the substrate. These result in a narrowing threshold voltage window after many cycles of programming and erasing. The major reliability problem of the MNOS transistor is the continuous loss of charge through the thin oxide. It should be pointed out that, unlike a floating-gate structure, the programming current has to pass through the entire channel region, so that the trapped charge is distributed uniformly throughout the channel. In a floating-gate transistor, the charge injected to the floating gate can redistribute itself within the gate material, and injection can take place locally anywhere along the channel.

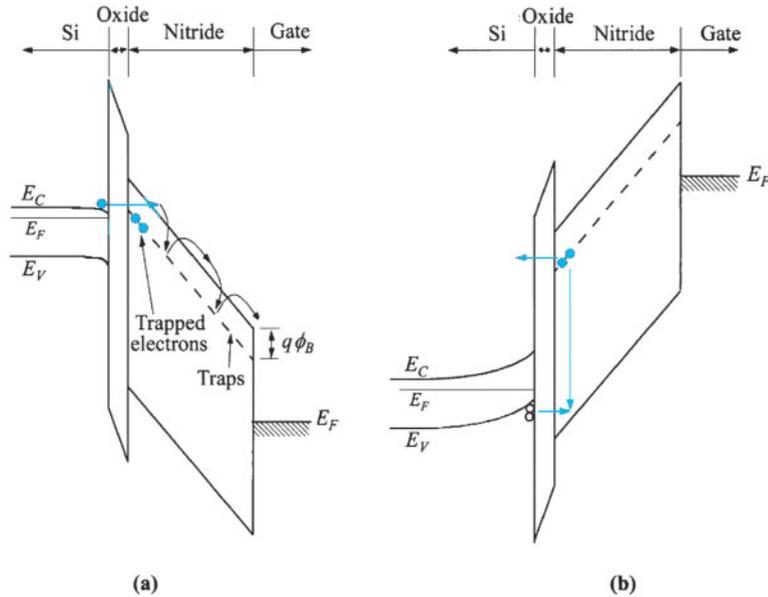


Fig. 33 Rewriting of MNOS memory. (a) Programming: electrons tunnel through the oxide and are trapped in the nitride. (b) Erasing: holes tunnel through the oxide to neutralize the trapped electrons and tunneling of trapped electrons.

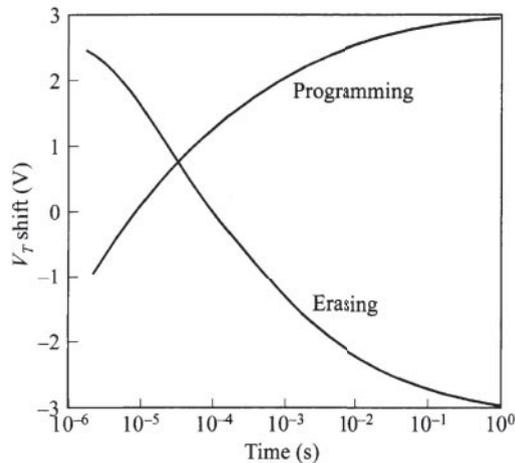


Fig. 34 Typical programming and erasing rates for MNOS transistor.

SONOS Transistor

The SONOS (silicon-oxide-nitride-oxide-silicon) transistor (Fig. 28d) is sometimes called the MONOS (metal-oxide-nitride-oxide-silicon) transistor. It is similar to an MNOS transistor except that it has an additional blocking oxide layer placed between the gate and the nitride layer, forming an ONO (oxide-nitride-oxide) stack. This top oxide layer is usually similar in thickness to the bottom oxide layer. The function of the blocking oxide is to prevent electron injection from the metal to the nitride layer during erase operation. As a result, a thinner nitride layer can be used, leading to lower programming voltage as well as better charge retention. The SONOS transistor now replaces the older MNOS configuration but the operation principle remains the same.

► 6.5 POWER MOSFET

The input impedance of MOS devices is very high because of the insulating SiO_2 between the gate and semiconductor channel. This feature makes the MOSFET an attractive candidate in power-device applications. Because of the high-input impedance, the gate leakage is very low, and thus the power MOSFET does not require complex input drive circuitry compared with bipolar devices. In addition, the switching speed of the power MOSFET is much faster than that of the power bipolar device. This is because the unipolar characteristics of MOS operation do not involve storage or recombination of minority carriers during turn-off.

The basic operation of power MOSFETs is the same as that of any MOSFET. However, the current handling capability is usually in the ampere range. Large current can be obtained with a large channel width. The drain to source blocking voltage is in the range of 50 to 100 volts or even higher. In general, power MOSFETs employ thicker oxides and deeper junctions, and have longer channel lengths. These generally post a penalty on device performance such as transconductance (g_m) and speed (f_T). However, power MOSFET applications have been on the rise, for example, due to the increasing demand of cellular phones and cellular base stations that require extra-high voltage.

Figure 35 shows three basic power MOSFET structures.¹⁸ Unlike the MOSFET structure used in advanced integrated circuits, the power MOSFETs employ a vertical structure with the source and drain at the top and bottom surfaces of the wafer, respectively. This vertical scheme has the benefit of large channel width and reduced field crowding at the gate. These properties are important for power applications.

Figure 35a is the V-MOSFET, in which the gate has a V-shaped groove. The V-shaped groove can be formed by preferential wet-etching using a KOH solution. When the gate voltage is larger than the threshold voltage, an inversion channel is induced at the surface along the edge of the V-shaped groove and forms the conductive path between source and drain. One main limitation of V-MOSFET development is related to process control. The high field at the tip of the V-shaped groove may lead to current crowding there and degrade device performance.

Figure 35b shows the cross-section of the U-MOSFET, which is similar to the V-MOSFET. The U-shaped trench is formed by reactive ion etching, and the electric fields at the bottom corners are substantially lower than that at the tip of the V-shaped groove. Another power MOSFET is the D-MOSFET, shown in Fig. 35c. The gate is formed at the top surface and then serves as a mask for the subsequent double-diffusion process. The double diffusion process (the reason why it is called “D”-MOSFET) is used to take advantage of the higher diffusion rate of the p -dopant (e.g., boron) than the n^+ dopant (e.g. phosphorus) to determine the channel length between the p -base and n^+ source portions. This technique can yield very short channels without depending on a lithographic mask. The advantages of D-MOSFET are its short drift time across the p -base region and the avoidance of high-field corners.

There is an n^- region in the drain for all the three power MOSFET structures. The doping concentration of the n^- drift region is lower than the p -base region. When a positive voltage is applied to the drain and the drain/ p -base junction is reverse biased, most of the depletion width will be developed across the n^- drift region. Consequently, the doping level and width of the n^- drift region are important parameters that determine the drain blocking voltage capability. On the other hand, there is a parasitic n - p - n^- - n^+ device in the power MOSFET structures. To prevent the action of the bipolar transistor during power MOSFET operation, the p -base and n^+ source (emitter) are shorted, as shown in Fig. 35. This can keep the p -base at a fixed potential.

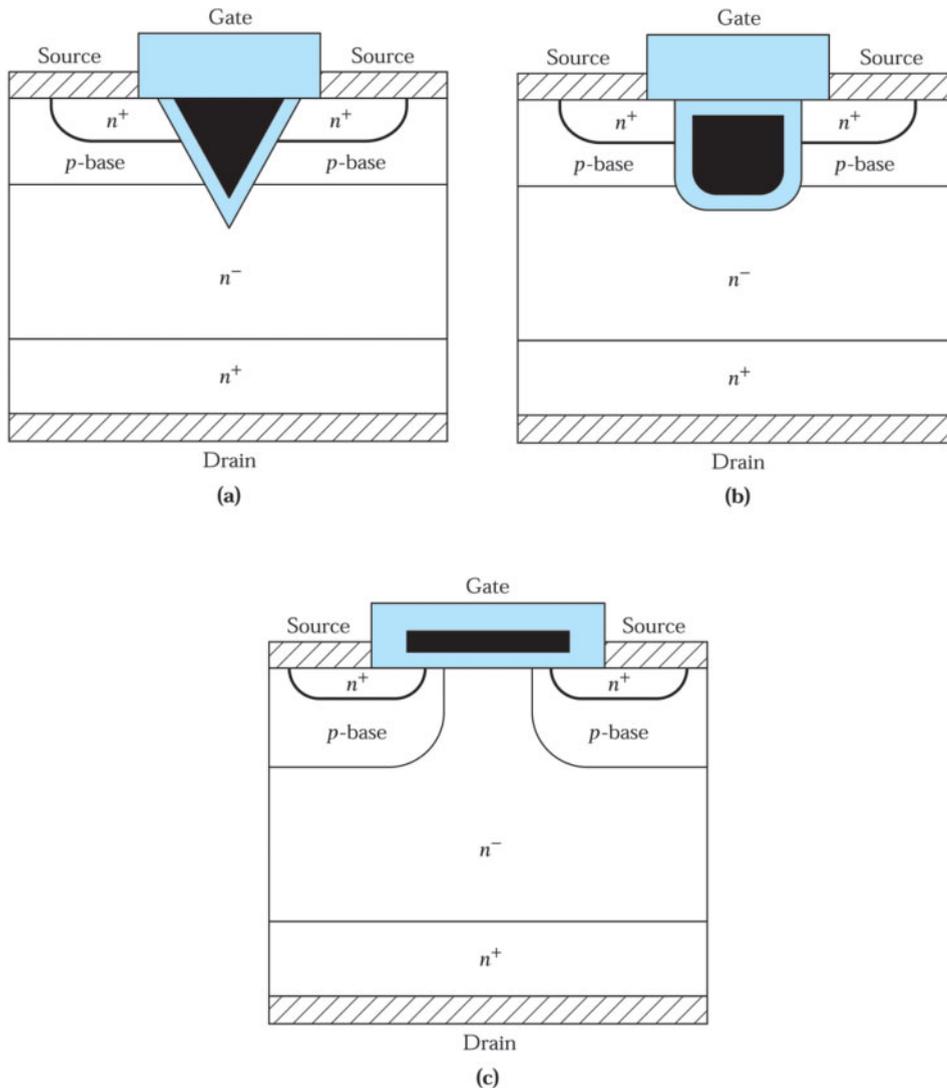


Fig. 35 (a) V-shaped MOS (VMOS), (b) U-shaped MOS (UMOS), and (c) double-diffused MOS (DMOS) power device structures.

► SUMMARY

The Si MOSFET is the most important device in advanced integrated-circuit (IC) applications. Its success is ascribed mainly to the high-quality SiO_2 material and its stable Si/SiO_2 interface properties. CMOS technology is currently the only viable solution meeting the stringent requirement for low power consumption in an IC chip, and is widely implemented. Superior power dissipation performance can be understood from the discussion of the CMOS inverter.

Scaling down of device dimensions is a continuing trend in CMOS technology to increase the device density, operating speed, and functionality of a chip. The short-channel effects, however, cause deviations in device operation and require attention in device scaling. Optimization of device structural parameters depends on the main requirement of the applications, such as minimized power consumption or faster operating speed,

TFT and SOI are MOSFET devices fabricated on insulating substrates, in contrast to the conventional MOSFETs fabricated on a bulk Si substrate. TFT uses an amorphous or polycrystalline semiconductor as the active channel layer. The carrier mobility of the TFT is degraded by the presence of a large number of defects in the channel. However, TFT can be applied to a large-area substrate, which is difficult for bulk MOS technology, e.g., the switching element of pixels in a large-area flat-panel display. TFT can also be used as the load devices in the SRAM cell. SOI MOSFETs use a monocrystalline Si channel layer. Compared with bulk-MOS devices, SOI devices provide lower parasitic junction capacitance and improved resistance to radiation damage. SOI is also more attractive for low-power, high-speed applications.

MOSFETs have been employed for semiconductor memory applications, including DRAM, SRAM, and nonvolatile memory. These products constitute a significant portion of the IC market. Owing to the aggressive shrinkage of device size, the storage capacity of the MOS memories improves rapidly. For example, the density of nonvolatile memories has doubled every 18 months and the single-electron memory has been projected to reach a multiterabit level. Finally, we considered three power MOSFETs. These devices use a vertical structure to allow higher operating voltage and current.

► REFERENCES

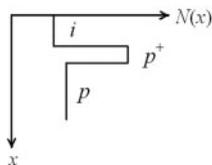
1. H. Kawaguchi, et al., "A Robust 0.15 μm CMOS Technology with CoSi₂ Salicide and Shallow Trench Isolation," in *Tech. Dig. Symp. VLSI Technol.*, p. 125 (1997).
2. L. D. Yau, "A Simple Theory to Predict the Threshold Voltage in Short-Channel IGFETs," *Solid-State Electron.*, 17, 1059 (1974).
3. R. H. Dennard, et al., "Design of Ion Implanted MOSFETs with Very Small Physical Dimensions," *IEEE J. Solid State Circuits*, SC-9, 256 (1974).
4. Y. Taur and T. K. Ning, *Physics of Modern VLSI Devices*, Cambridge Univ. Press, London, 1998.
5. H-S. P. Wong, "MOSFET Fundamentals," in *ULSI Devices*, C. Y. Chang and S. M. Sze, Eds., Wiley Interscience, New York, 1999. Fu-Liang Yang et al., "5nm-Gate Nanowire FinFET," in *Tech. Dig. Symp. VLSI Technol.*, p. 196 (2004).
7. R. R. Troutman, *Latch-up in CMOS Technology*, Kluwer, Boston, 1986.
8. A. El Gamal and H. Eltoukhy, "CMOS Image Sensors," *IEEE Circuits Dev. Mag.*, 6, 2005.
9. A. Theuwissen, "CMOS Image Sensors: State-Of-The-Art and Future Perspectives," *Proc. 37th Eur. Solid State Device Res. Conf.*, 21, 2007.
10. K. K. Ng, *Complete Guide to Semiconductor Devices*, 2nd Ed., Wiley/IEEE Press, Hoboken, New Jersey, 2002.
11. J. P. Colinge, *Silicon-on-Insulator Technology: Materials to VLSI*, Kluwer, Boston, 1991. (p. 185, 186, 187)
12. B. S. Doyle, S. Datta, M. Doczy, S. Harelend, B. Jin, J. Kavalieros, T. Linton, A. Murthy, R. Rios, and R. Chau, "High Performance Fully-Depleted Tri-Gate CMOS Transistors," *IEEE Electron Dev. Lett.*, EDL-24, 263 (2003).
13. J. M. Hergenrother, G. D. Wilk, T. Nigam, F. P. Klemens, D. Monroe, P. J. Silverman, T. W. Sorsch, B. Busch, M. L. Green, M. R. Baker et al., "50 nm Vertical Replacement-Gate (VRG) nMOSFETs with ALD HfO₂ and Al₂O₃ Gate Dielectrics," *Tech. Dig. IEEE IEDM*, p.51, 2001.
14. (a) R. H. Dennard, "Field-effect Transistor Memory," U.S. Patent 3,387,286. (b) R. H. Dennard, "Evolution of the MOSFET DRAM-A Personal View," *IEEE Trans. Electron Devices*, ED-31, 1549 (1984).

15. D. Kahng and S. M. Sze, "A Floating Gate and Its Application to Memory Devices," *Bell System Tech. J.*, 46,1288 (1967).
16. F. Masuoka, M. Asano, H. Iwahashi, T. Komuro, and S. Tanaka, "A New Flash E²PROM Cell Using Triple Polysilicon Technology," *IEEE Tech. Dig. Int. Electron. Devices Meet.*, p.464, 1984.
17. S. M. Sze, "Evolution of Nonvolatile Semiconductor Memory; from Floating-Gate Concept to Single-Electron Memory Cell," in S. Luryi, J. Xu, and A. Zaslavsky, Eds., *Future Trends in Microelectronics*, Wiley Interscience, New York, 1999.
18. B. J. Baliga, *Power Semiconductor Devices*, PWS Publishers, Boston, 1996.

► **PROBLEMS (* DENOTES DIFFICULT PROBLEMS)**

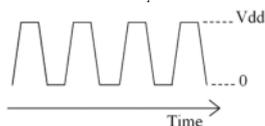
FOR SECTION 6.1 MOSFET SCALING

1. When the linear dimensions of MOSFET are scaled down by a factor of 10 based on the constant field scaling, (a) what is the corresponding factor for the corresponding switching energy? (b) what is the scaled power-delay product, assuming the product is 1 J for the original large device?
- *2. Based on the charge-sharing model, Fig. 2c, show that the threshold voltage roll-off is given by Eq. 1.
3. Assume the retrograde channel doping profile has a peak level slightly below the semiconductor surface shown in the figure below is used in an *n*-channel MOSFET with *n*⁺ polysilicon gate. Draw the energy band diagram of the MOSFET from the gate to substrate, when the gate is biased at threshold voltage.



FOR SECTION 6.2 CMOS AND BiCMOS

4. Describe the pros and cons of BiCMOS.
5. For the CMOS inverter shown in Fig. 14, if a voltage train (V_{in}) is applied to the input terminal of a CMOS inverter: (a) Schematically plot the corresponding output voltage (V_{out}). (b) What states of NMOS and PMOS correspond to points labeled A, B, C, D, and point between C and D? (c) If the V_{tn} of *n*-channel and *p*-channel MOSFETs are V_{tn} and V_{tp} , indicate the point when NMOS just becoming saturated from the linear region with the following parameters.
For *n*-channel MOSFET: $\mu_{ns}C_o(Z/L) = 20 \text{ mA/V}^2$ and $V_{tn} = 2\text{V}$.
For *p*-channel MOSFET: $\mu_{ps}C_o(Z/L) = 20 \text{ mA/V}^2$ and $V_{tp} = 1\text{V}$.



FOR SECTION 6.3 MOSFET ON INSULATOR

6. For an *n*-channel FD-SOI device having $N_A = 5 \times 10^{17} \text{ cm}^{-3}$ and $d = 4 \text{ nm}$, calculate the maximum allowable thickness for Si channel layer (d_{Si}).

7. For an n -channel SOI device with n^+ -polysilicon gate having $N_A = 5 \times 10^{17} \text{ cm}^{-3}$, $d = 4 \text{ nm}$, and $d_{Si} = 30 \text{ nm}$, calculate the threshold voltage. Assume that Q_f , Q_{ov} , and Q_m are all zero.
8. For the device in Prob. 6, calculate the range of V_T distribution if the thickness variation of d_{Si} across the wafer is $\pm 5 \text{ nm}$.
9. What is the advantage of using FinFET for CMOS applications compared with other planar MOSFETs in CMOS?

FOR SECTION 6.4 MOS MEMORY STRUCTURES

10. What is the capacitance of a DRAM capacitor if it is planar, $1 \mu\text{m} \times 1 \mu\text{m}$, with an oxide thickness of 10 nm ? Calculate the capacitance if the same surface area is used for a trench that is $7 \mu\text{m}$ deep and has the same oxide thickness.
11. For DRAM operation, assume that we need a minimum of 10^5 electrons for the MOS storage capacitor. If the capacitor has an area of $0.5 \mu\text{m} \times 0.5 \mu\text{m}$ on the wafer surface, an oxide thickness of 5 nm , and is fully charged to 2 V , what is the required minimum depth of a rectangular-trench capacitor?
12. A DRAM must operate with a minimum refresh time of 4 ms . The storage capacitor in each cell has a capacitance of 50 fF and is fully charged to 5 V . Find the worst-case leakage current (i.e., during the refresh cycle 50% of the stored charge is lost) that the dynamic node can tolerate.
13. A floating-gate nonvolatile memory has an initial threshold voltage of -2 V , and a linear-region drain conductance of $10 \mu\text{mhos}$ at a gate voltage of -5 V . After a write operation, the drain conductance increases to $40 \mu\text{mhos}$ at the same gate voltage. Find the threshold voltage shift.
14. For a floating-gate nonvolatile memory device, the lower insulator has a dielectric constant of 4 and is 10 nm thick. The insulator above the floating gate has a dielectric constant of 10 and is 100 nm thick. If the current density $J_1 = \sigma E_1$ where $\sigma = 10^{-7} \text{ S/cm}$, and the current in the upper insulator is zero, find the threshold voltage shift for a sufficiently long time that J_1 becomes negligibly small. The applied voltage on the control gate is 10 V .
15. A floating-gate nonvolatile semiconductor memory has a total capacitance of 3.71 fF , a control gate to floating-gate capacitance of 2.59 fF , a drain to floating gate capacitance of 0.49 fF , and a floating -gate to substrate capacitance of 0.14 fF . How many electrons are needed to shift the measured threshold by 0.5 V (measured from the control gate)?
16. Fill in the following table with simple description of its characteristics.

| | Cell size | Write one byte rate | Rewrite cycle | Keep data without power | Applications |
|-------|-----------|---------------------|---------------|-------------------------|--------------|
| SRAM | | | | | |
| DRAM | | | | | |
| Flash | | | | | |

FOR SECTION 6.5 POWER MOSFET

17. A power MOSFET has an n^+ -polysilicon gate and a p -base with $N_A = 10^{17} \text{ cm}^{-3}$. Gate oxide thickness $d = 100 \text{ nm}$. Calculate the threshold voltage.
18. For the device in Prob. 16, calculate the effect of a positive fixed charge density of $5 \times 10^{11} \text{ cm}^{-3}$ on the threshold voltage.

MESFET and Related Devices

- ▶ 7.1 METAL-SEMICONDUCTOR CONTACTS
 - ▶ 7.2 MESFET
 - ▶ 7.3 MODFET
 - ▶ SUMMARY
-

The metal-semiconductor field-effect transistor (MESFET) has current-voltage characteristics similar to those of a metal-oxide-semiconductor field-effect transistor (MOSFET). However, it uses a metal-semiconductor rectifying contact instead of a MOS structure for the gate electrode. In addition, the source and drain contacts of MESFET are ohmic,* whereas in a MOSFET they are p - n junctions.

Like other field-effect devices, a MESFET has a negative temperature coefficient at high current levels; that is, the current decreases as temperature increases. This characteristic leads to more uniform temperature distribution and the device is therefore thermally stable, even when the active area is large or when many devices are connected in parallel. Furthermore, because MESFETs can be made from compound semiconductors with high electron mobilities, such as GaAs and InP, they have higher switching speeds and higher cutoff frequencies than do silicon MOSFETs.

The basic building block of a MESFET is the metal-semiconductor contact. This contact is electrically similar to a one-sided abrupt p - n junction, yet it can be operated as a majority carrier device with inherently fast response. There are two types of metal-semiconductor contacts: the rectifying and the nonrectifying or ohmic types. In this chapter, we begin with the two types of contacts and then consider the basic characteristics and microwave performance of the MESFET. In the last section we discuss modulation-doped field-effect transistor (MODFET), which is similar to the device configuration of a MESFET but can offer even higher-speed performance.

Specifically, we cover the following topics:

- The rectifying metal-semiconductor contact and its current-voltage characteristics.
- The ohmic metal-semiconductor contact and its specific contact resistance.
- The MESFET and its high-frequency performance.
- The MODFET and its two-dimensional electron gas.
- A comparison of three field-effect transistors—MOSFET, MESFET, and MODFET.

*The concept of rectifying was discussed in Chapter 3; the concept of ohmic contact is presented in Section 7.1.

► 7.1 METAL-SEMICONDUCTOR CONTACTS

The first practical semiconductor device was the metal-semiconductor contact in the form of a point-contact rectifier, that is, a metallic whisker pressed against a semiconductor. The device found many applications beginning in 1904. In 1938, Schottky suggested that the rectifying behavior could arise from a potential barrier as a result of the stable space charges in the semiconductor.¹ The model arising from this concept is known as the Schottky barrier. Metal-semiconductor contacts can also be nonrectifying; that is, the contact has negligible resistance regardless of the polarity of the applied voltage. This type of contact is called an ohmic contact. All semiconductor devices as well as integrated circuits need ohmic contact to make connections to other devices in an electronic system. We consider the energy band diagram and the current-voltage characteristics of both the rectifying and ohmic metal-semiconductor contacts.

7.1.1 Basic Characteristics

The characteristics of point-contact rectifiers were not reproducible from one device to another. The contact was just a simple mechanical contact or formed by an electrical discharge process that could result in a small alloyed p - n junction. The advantage of a point-contact rectifier is its small area, which can give very small capacitance, a desirable feature for microwave application. The rectifiers are subject to wide variations such as the whisker pressure, contact area, crystal structure, whisker composition, and heat or forming processes, and they have been largely replaced by metal-semiconductor contacts fabricated by planar processes (see Chapters 11–15). A schematic diagram of such a device is shown in Fig. 1*a*. To fabricate the device, a window is opened in an oxide layer, and a metal layer is deposited in a vacuum system. The metal layer covering the window is subsequently defined by a lithographic step. We consider the one-dimensional structure of the metal-semiconductor contact shown in Fig. 1*b*, which corresponds to the central section in Fig. 1*a*, between the dashed lines.

Figure 2*a* shows the energy band diagram of an isolated metal adjacent to an isolated n -type semiconductor. Note that the metal work function $q\phi_m$ is generally different from the semiconductor work function $q\phi_s$. The work function is defined as the energy difference between the Fermi level and the vacuum level. Also shown is the electron affinity $q\chi$, which is the energy difference between the conduction band edge and the vacuum

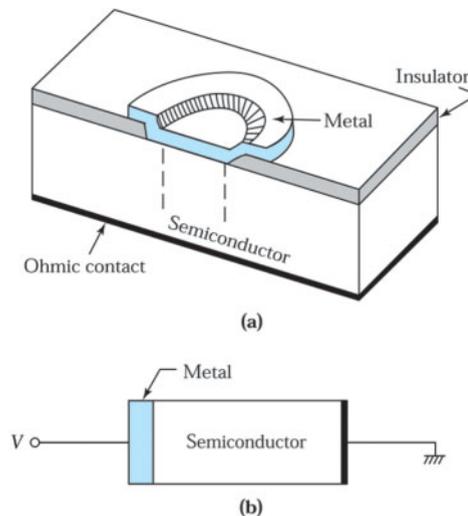


Fig. 1 (a) Perspective view of a metal-semiconductor contact fabricated by the planar process. (b) One-dimensional structure of a metal-semiconductor contact.

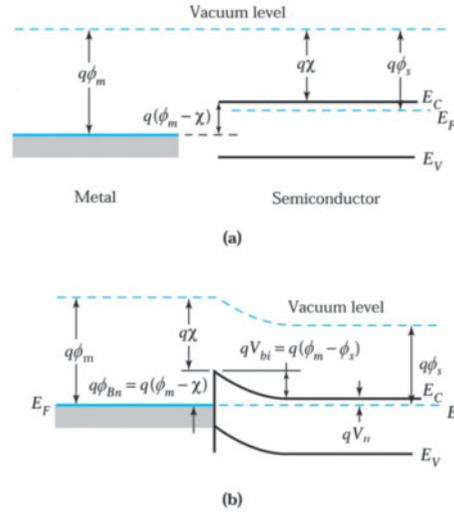


Fig. 2 (a) Energy band diagram of an isolated metal adjacent to an isolated n-type semiconductor under thermal nonequilibrium condition. (b) Energy band diagram of a metal-semiconductor contact in thermal equilibrium.

level in the semiconductor. When the metal makes intimate contact with the semiconductor, the Fermi levels in the two materials must be equal at thermal equilibrium. In addition, the vacuum level must be continuous. These two requirements determine a unique energy band diagram for the ideal metal-semiconductor contact, as shown in Fig. 2b.

For this ideal case, the barrier height $q\phi_{Bn}$ is simply the difference between the metal work function and the semiconductor electron affinity[§]:

$$q\phi_{Bn} = q\phi_m - q\chi. \quad (1)$$

Similarly, for the case of an ideal contact between a metal and a *p*-type semiconductor, the barrier height $q\phi_{Bp}$ is given by

$$q\phi_{Bp} = E_g - (q\phi_m - q\chi), \quad (2)$$

where E_g is the bandgap of the semiconductor. Therefore, for a given semiconductor and for any metal, the sum of the barrier heights on *n*-type and *p*-type substrates is expected to be equal to the bandgap:

$$\boxed{q(\phi_{Bn} + \phi_{Bp}) = E_g.} \quad (3)$$

On the semiconductor side in Fig. 2b, V_{bi} is the built-in potential that is seen by electrons in the conduction band trying to move into the metal.

$$V_{bi} = \phi_{Bn} - V_n. \quad (4)$$

The qV_n is the distance between the bottom of the conduction band and the Fermi level. Similar results can be given for the *p*-type semiconductor.

[§] Both $q\phi_{Bn}$ (in electron volts) and ϕ_{Bn} (in volts) are referred to as the barrier height.

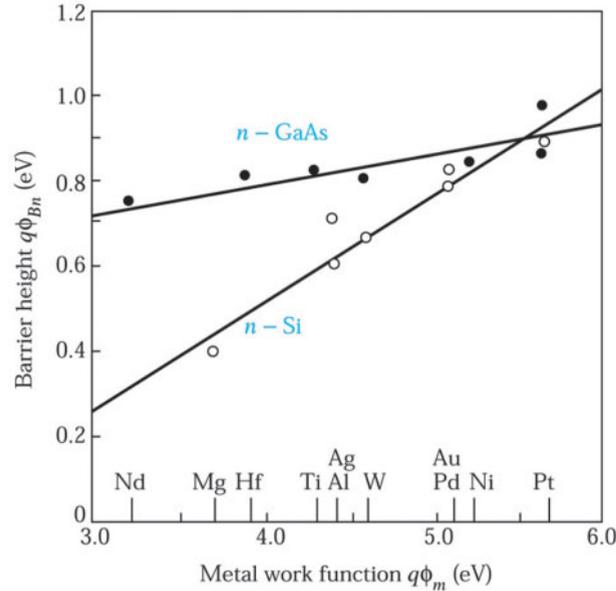


Fig. 3 Measured barrier height for metal-silicon and metal-gallium arsenide contacts.^{2,3}

Figure 3 shows the measured barrier heights for n -type silicon² and n -type gallium arsenide.³ Note that $q\phi_{Bn}$ increases with increasing $q\phi_m$. However, the dependence is not as strong as predicted by Eq. 1. This is because in practical Schottky diodes, the disruption of the crystal lattice at the semiconductor surface produces a large number of surface energy states located in the forbidden bandgap. These surface states can act as donors or acceptors that influence the final barrier height. For silicon and gallium arsenide, Eq. 1 generally underestimates the n -type barrier height and Eq. 2 overestimates the p -type barrier height. The sum of $q\phi_{Bn}$ and $q\phi_{Bp}$, however, is in agreement with Eq. 3.

Figure 4 shows energy band diagrams for metals on both n -type and p -type semiconductors under different biasing conditions. Consider the n -type semiconductor first. When the bias voltage is zero, as shown in the left side of Fig. 4a, the band diagram is under a thermal equilibrium condition. The Fermi levels for both materials are equal. If we apply a positive voltage to the metal with respect to the n -type semiconductor, the semiconductor-to-metal built-in potential decreases as shown on the left side of Fig. 4b. This is a forward bias. When a forward bias is applied, electrons can move easily from the semiconductor into the metal because the barrier has been reduced by a voltage V_F . For a reverse bias (i.e., a negative voltage is applied to the metal), the barrier is increased by a voltage V_R , as depicted on the left side of Fig. 4c. It is more difficult for electrons to flow from the semiconductor into the metal. We have similar results for p -type semiconductor, however, the polarities must be reversed. In the following derivations, we consider only the metal- n -type semiconductor contact. The results are equally applicable to a p -type semiconductor with an appropriate change of polarities.

The charge and field distributions for a metal-semiconductor contact are shown in Fig. 5a and 5b, respectively. The metal is assumed to be a perfect conductor; the charge transferred to it from the semiconductor exists in a very narrow region at the metal surface. The extent of the space charge in the semiconductor is W , i.e., $\rho_s = qN_D$ for $x < W$ and $\rho_s = 0$ for $x > W$. Thus, the charge distribution is identical to that of a one-sided abrupt p^+-n junction.

The magnitude of the electric field decreases linearly with distance. The maximum electric field \mathcal{E}_m is located at the interface. The electric field distribution is then given by

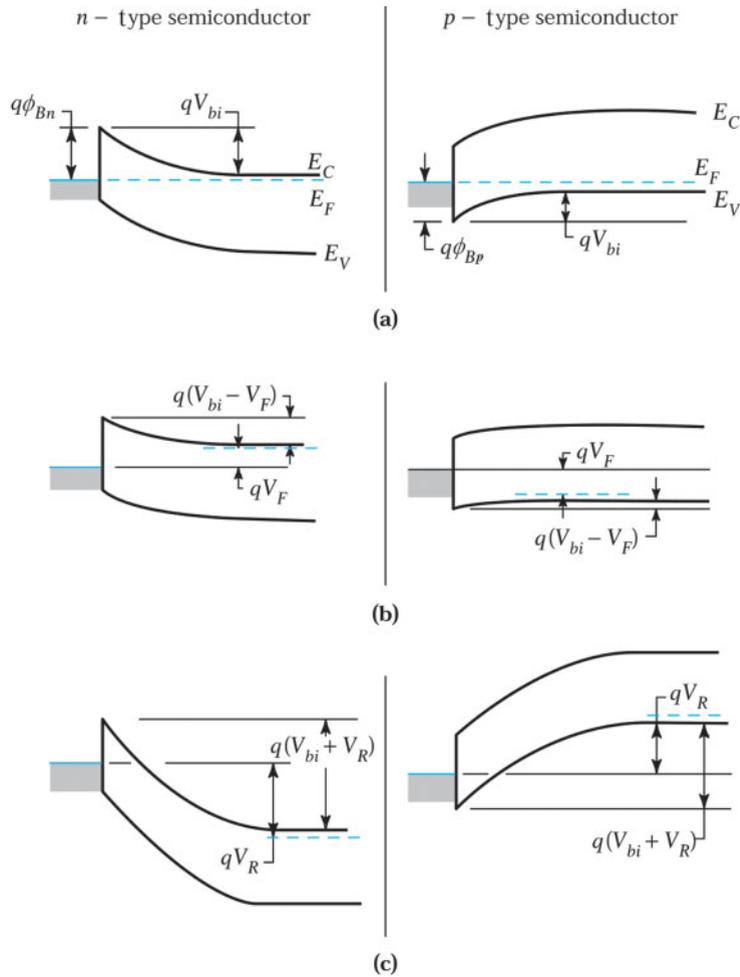


Fig. 4 Energy band diagrams of metal n-type and p-type semiconductors under different biasing conditions: (a) thermal equilibrium; (b) forward bias; and (c) reverse bias.

$$|\mathcal{E}(x)| = \frac{qN_D}{\epsilon_s}(W - x) = \mathcal{E}_m - \frac{qN_D}{\epsilon_s}x, \quad (5)$$

$$\mathcal{E}_m = \frac{qN_D W}{\epsilon_s}, \quad (6)$$

where ϵ_s is the dielectric permittivity of the semiconductor. The voltage across the space-charge region, which is represented by the area under the field curve in the Fig. 5b, is given by

$$V_{bi} - V = \frac{\mathcal{E}_m W}{2} = \frac{qN_D W^2}{2\epsilon_s}. \quad (7)$$

The depletion-layer width W is expressed as

$$W = \sqrt{2\epsilon_s(V_{bi} - V)/qN_D}, \quad (8)$$

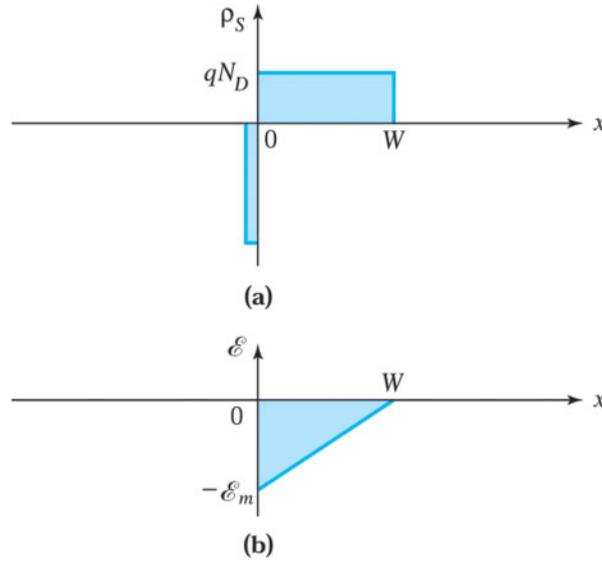


Fig. 5 (a) Charge distribution and (b) electric-field distribution in a metal-semiconductor contact.

and the space-charge density, Q_{sc} , in the semiconductor is given as

$$Q_{sc} = qN_D W = \sqrt{2q\epsilon_s N_D (V_{bi} - V)} \text{ C/cm}^2, \quad (9)$$

where the voltage V equal to $+V_f$ for forward bias and to $-V_r$ for reverse bias. The depletion-layer capacitance C per unit area can be calculated by using Eq. 9:

$$C = \left| \frac{\partial Q_{sc}}{\partial V} \right| = \sqrt{\frac{q\epsilon_s N_D}{2(V_{bi} - V)}} = \frac{\epsilon_s}{W} \text{ F/cm}^2 \quad (10)$$

and

$$\boxed{\frac{1}{C^2} = \frac{2(V_{bi} - V)}{q\epsilon_s N_D} \text{ (F/cm}^2\text{)}^{-2}}. \quad (11)$$

We can differentiate $1/C^2$ with respect to V . Rearranging terms we obtain:

$$N_D = \frac{2}{q\epsilon_s} \left[\frac{-1}{d(1/C^2)/dV} \right]. \quad (12)$$

Thus, measurements of the capacitance C per unit area as a function of voltage can provide the impurity distribution from Eq. 12. If N_D is constant throughout the depletion region, we should obtain a straight line by plotting $1/C^2$ versus V . Figure 6 is a plot of the measured capacitance versus voltage for tungsten-silicon and tungsten-gallium arsenide Schottky diodes.⁴ From Eq. 11, the intercept at $1/C^2 = 0$ corresponds to the built-in potential V_{bi} . Once V_{bi} is determined, the barrier height ϕ_{Bn} can be calculated from Eq. 4.

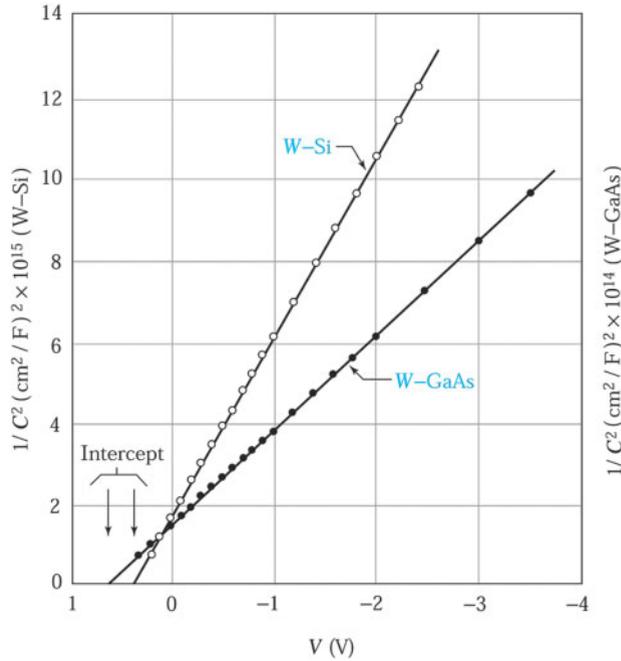


Fig. 6 $1/C^2$ versus applied voltage for W-Si and W-GaAs diodes.⁴

► EXAMPLE 1

Find the donor concentration and the barrier height of the tungsten-silicon Schottky diode shown in Fig. 6.

SOLUTION The plot of $1/C^2$ versus V is a straight line, which implies that the donor concentration is constant throughout the depletion region. We find

$$\frac{d(1/C^2)}{dV} = \frac{6.2 \times 10^{15} - 1.8 \times 10^{15}}{-1 - 0} = -4.4 \times 10^{15} \frac{(\text{cm}^2/\text{F})^2}{\text{V}}.$$

From Eq. 12,

$$N_D = \left[\frac{2}{1.6 \times 10^{-19} \times (11.9 \times 8.85 \times 10^{-14})} \right] \times \left(\frac{1}{4.4 \times 10^{15}} \right) = 2.7 \times 10^{15} \text{ cm}^{-3},$$

$$V_n = 0.0259 \times \ln \left(\frac{2.86 \times 10^{19}}{2.7 \times 10^{15}} \right) = 0.24 \text{ V}.$$

Since the intercept V_{bi} is 0.42 V, then the barrier height is $\phi_{Bn} = 0.42 + 0.24 = 0.66 \text{ V}$. ◀

7.1.2 The Schottky Barrier

A Schottky barrier refers to a metal-semiconductor contact having a large barrier height (i.e., ϕ_{Bn} or $\phi_{Bp} \gg kT$) and a low doping concentration that is less than the density of states in the conduction band or valence band.

The current transport in a Schottky barrier is due mainly to majority carriers, unlike a p - n junction where current transport is due mainly to minority carriers. For Schottky diodes operated at moderate temperature (e.g., 300 K), the dominant transport mechanism is thermionic emission of majority carriers from the semiconductor over the potential barrier into the metal.

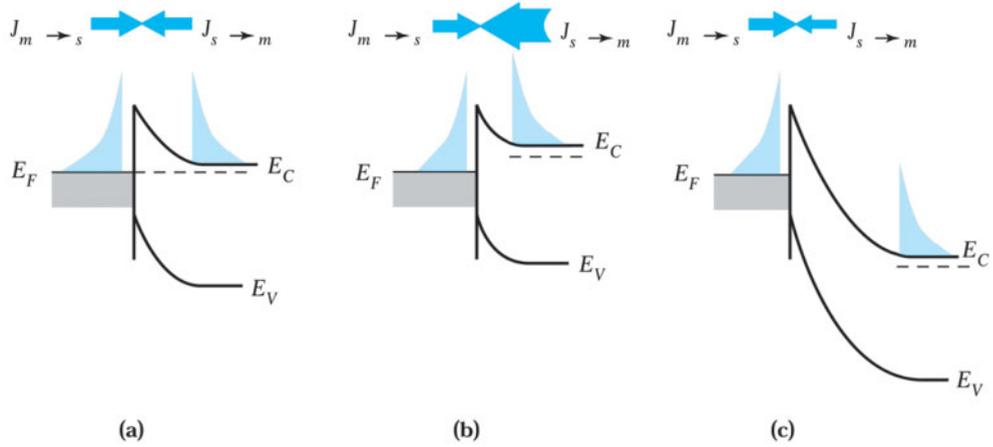


Fig. 7 Current transport by the thermionic emission process. (a) Thermal equilibrium; (b) forward bias; and (c) reverse bias.⁵

Figure 7 illustrates the thermionic emission process.⁵ At thermal equilibrium (Fig. 7a), the current density is balanced by two equal and opposite flows of carriers, and thus there is zero net current. Electrons in the semiconductor tend to flow (or emit) into the metal, and there is an opposing balanced flow of electrons from metal into the semiconductor. These current components are proportional to the density of electrons at the boundary.

As discussed in Section 2.5 of Chapter 2, at the semiconductor surface an electron can be thermionically emitted into the metal if its energy is above the barrier height. Here the semiconductor work function $q\phi_s$ is replaced by $q\phi_{Bn}$, and

$$n_{th} = N_C \exp\left(-\frac{q\phi_{Bn}}{kT}\right), \quad (13)$$

where N_C is the density of states in the conduction band. At thermal equilibrium we have

$$|J_{m \rightarrow s}| = |J_{s \rightarrow m}| \propto n_{th} \quad (14)$$

or

$$|J_{m \rightarrow s}| = |J_{s \rightarrow m}| = C_1 N_C \exp\left(-\frac{q\phi_{Bn}}{kT}\right), \quad (14a)$$

where $J_{m \rightarrow s}$ is the current from the metal to the semiconductor, $J_{s \rightarrow m}$ is the current from the semiconductor to the metal, and C_1 is a proportionality constant.

When a forward bias V_F is applied to the contact (Fig. 7b), the electrostatic potential difference across the barrier is reduced, and the electron density at the surface increases to

$$n_{th} = N_C \exp\left[-\frac{q(\phi_{Bn} - V_F)}{kT}\right]. \quad (15)$$

The current $J_{s \rightarrow m}$ that results from the electron flow out of the semiconductor is therefore altered by the same factor (Fig. 7b). The flux of electrons from the metal to the semiconductor, however, remains the same because the barrier ϕ_{Bn} remains at its equilibrium value. The net current under forward bias is then

$$\begin{aligned}
J &= J_{s \rightarrow m} - J_{m \rightarrow s} \\
&= C_1 N_C \exp\left[-\frac{q(\phi_{Bn} - V_F)}{kT}\right] - C_1 N_C \exp\left(-\frac{q\phi_{Bn}}{kT}\right) \\
&= C_1 N_C e^{-q\phi_{Bn}/kT} \left(e^{qV_F/kT} - 1\right).
\end{aligned} \tag{16}$$

With the same argument for the reverse-bias condition (see Fig. 7c), the expression for the net current is identical to Eq. 16 except that V_F is replaced by $-V_R$.

The coefficient $C_1 N_C$ is found to be equal to $A^* T^2$, where A^* is called the *effective Richardson constant* (in units of $\text{A}/\text{K}^2\text{-cm}^2$) and T is the absolute temperature. The value of A^* depends on the effective mass and is equal to 110 and 32 for n - and p -type silicon, respectively, and 8 and 74 for n - and p -type gallium arsenide, respectively.⁶

The current-voltage characteristic of a metal-semiconductor contact under thermionic emission condition is then

$$J = J_s \left(e^{qV/kT} - 1 \right), \tag{17}$$

$$J_s = A^* T^2 e^{-q\phi_{Bn}/kT}, \tag{17a}$$

where J_s is the saturation current density and the applied voltage V is positive for forward bias and negative for reverse bias. Experimental forward I - V characteristics of two Schottky diodes⁴ are shown in Fig. 8. By extrapolating the forward I - V curve to $V = 0$, we can find J_s . From J_s and Eq. 17a we can obtain the barrier height.

In addition to the majority carrier (electron) current, a minority-carrier (hole) current exists in a metal n -type semiconductor contact. The electron-hole pairs can be easily created in the valence band (interband transition) in the depletion region. Electrons in the valence band flow into the metal because there is no barrier there and holes diffuse into the semiconductor to form the minority current under forward bias. The hole diffusion current is the same as in a p - n junction, which is described in Chapter 3. The current density is given by

$$J_p = J_{po} \left(e^{qV/kT} - 1 \right), \tag{18}$$

where

$$J_{po} = \frac{qD_p n_i^2}{L_p N_D}. \tag{18a}$$

Under normal operating conditions, the minority-carrier diffusion current is orders of magnitude smaller than the majority-carrier current. Therefore, a Schottky diode is a unipolar device (i.e., predominately only one type of carrier participates in the conduction process). The minimum minority-carrier storage makes the Schottky barrier to operate at much higher frequencies (~ 100 GHz) compared to a p - n junction (~ 1 GHz).

► EXAMPLE 2

For a tungsten-silicon Schottky diode with $N_D = 10^{16} \text{ cm}^{-3}$, find the barrier height and depletion-layer width from Fig. 8. Compare the saturation current J_s with J_{po} , assuming that the minority-carrier lifetime in Si is 10^{-6} s.

SOLUTION From Fig. 8, we have $J_s = 6.5 \times 10^{-5} \text{ A/cm}^2$. The barrier height can be obtained from Eq. 17a:

$$\phi_{Bn} = 0.0259 \times \ln\left(\frac{110 \times 300^2}{6.5 \times 10^{-5}}\right) = 0.67 \text{ V}.$$

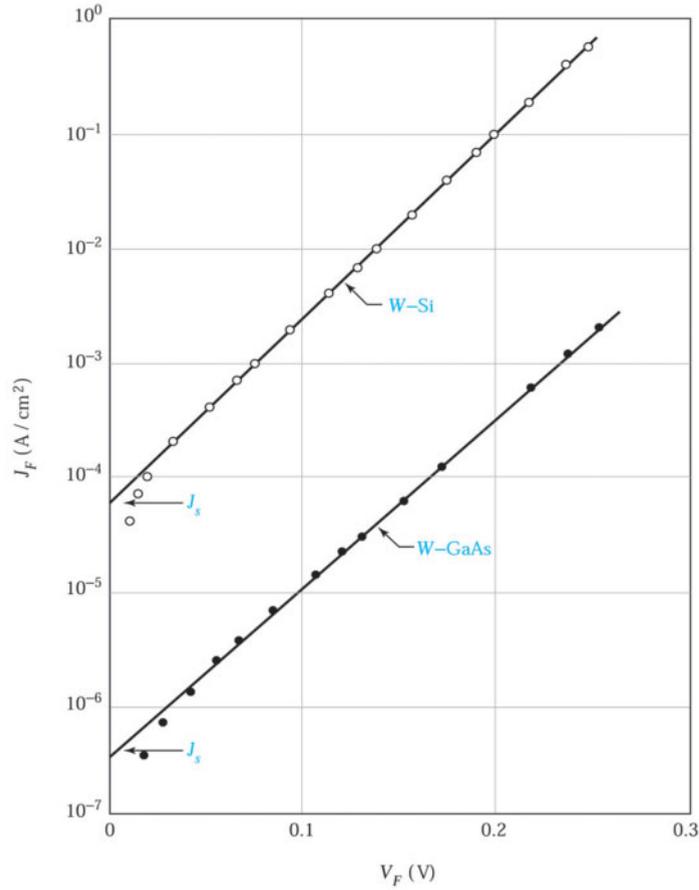


Fig. 8 Forward current density versus applied voltage of W-Si and W-GaAs diodes.⁴

This result is in close agreement with the C - V measurement (see Fig. 6 and Ex. 1). The built-in potential is given by $\phi_{Bn} - V_n$, where

$$V_n = 0.0259 \times \ln\left(\frac{N_C}{N_D}\right) = 0.0259 \ln\left(\frac{2.86 \times 10^{19}}{1 \times 10^{16}}\right) = 0.17 \text{ V.}$$

Therefore

$$V_{bi} = 0.67 - 0.17 = 0.50 \text{ V.}$$

The depletion-layer width at thermal equilibrium is given by Eq. 8 with $V = 0$:

$$W = \sqrt{\frac{2\epsilon_s V_{bi}}{qN_D}} = 2.6 \times 10^{-5} \text{ cm.}$$

To calculate the minority-carrier current density J_{po} , we need to know D_p , which is $10 \text{ cm}^2/\text{s}$ for $N_D = 10^{16} \text{ cm}^{-3}$, and L_p which is $\sqrt{D_p \tau_p} = \sqrt{10 \times 10^{-6}} = 3.1 \times 10^{-3} \text{ cm}$. Therefore,

$$J_{po} = \frac{qD_p n_i^2}{L_p N_D} = \frac{1.6 \times 10^{-19} \times 10 \times (9.65 \times 10^9)^2}{(3.1 \times 10^{-3}) \times 10^{16}} = 4.8 \times 10^{-12} \text{ A/cm}^2.$$

The ratio of the two current densities is

$$\frac{J_s}{J_{po}} = \frac{6.5 \times 10^{-5}}{4.8 \times 10^{-12}} = 1.3 \times 10^7.$$

From the comparison, we see that the majority-carrier current is over seven orders of magnitude greater than the minority-carrier current. ◀

7.1.3 The Ohmic Contact

An ohmic contact is defined as a metal-semiconductor contact that has a negligible contact resistance relative to the bulk or series resistance of the semiconductor. A satisfactory ohmic contact should not significantly degrade device performance and can pass the required current with a voltage drop that is small compared with the drop across the active region of the device.

A figure of merit for ohmic contacts is the specific contact resistance R_C , defined as

$$R_C \equiv \left(\frac{\partial J}{\partial V} \right)_{V=0}^{-1} \Omega \text{-cm}^2. \quad (19)$$

For metal-semiconductor contacts with low doping concentrations, the thermionic-emission current dominates the current transport, as given by Eq. 17. Therefore,

$$R_C = \frac{k}{qA^*T} \exp\left(\frac{q\phi_{Bn}}{kT}\right). \quad (20)$$

Equation 20 shows that a metal-semiconductor contact with a low barrier height should be used to obtain a small R_C .

For contacts with high doping concentration, the barrier width becomes very narrow, and the tunneling current becomes dominant. The tunneling current, as described in the upper inset of Fig. 9, is proportional to the tunneling probability, which is given in Section 2.6 of Chapter 2:

$$I \sim \exp\left[-2W\sqrt{2m_n(q\phi_{Bn} - qV)/\hbar^2}\right], \quad (21)$$

where W is the depletion-layer width, which can be approximated as $\sqrt{(2\varepsilon_s/qN_D)(\phi_{Bn} - V)}$, m_n is the effective mass, and \hbar is the reduced Planck constant. Substituting W into Eq. 21, we obtain

$$I \sim \exp\left[-\frac{C_2(\phi_{Bn} - V)}{\sqrt{N_D}}\right], \quad (22)$$

where $C_2 = 4\sqrt{m_n\varepsilon_s}/\hbar$. The specific contact resistance for high doping is thus

$$R_C \sim \exp\left(\frac{C_2\phi_{Bn}}{\sqrt{N_D}}\right) = \exp\left(\frac{4\sqrt{m_n\varepsilon_s}\phi_{Bn}}{\sqrt{N_D}\hbar}\right). \quad (23)$$

Equation 23 shows that in the tunneling range the specific contact resistance depends strongly on doping concentration and varies exponentially with the factor $\phi_{Bn} / \sqrt{N_D}$.

The calculated values of R_C are plotted⁶ in Fig. 9 as a function of $1/\sqrt{N_D}$. For $N_D \geq 10^{19} \text{ cm}^{-3}$, R_C is dominated by the tunneling process and decreases rapidly with increased doping. On the other hand, for $N_D \leq 10^{17} \text{ cm}^{-3}$, the current is due to thermionic emission, and R_C is essentially independent of doping. Also shown in Fig. 9 are experimental data for platinum silicide-silicon (PtSi-Si) and aluminum-silicon (Al-Si) diodes. They are in close agreement with the calculated values. Figure 9 shows that a high doping concentration, a low barrier height, or both must be used to obtain a low value of R_C . These two approaches are used for all practical ohmic contacts.

► EXAMPLE 3

An ohmic contact has an area of 10^{-5} cm^2 and a specific contact resistance of $10^{-6} \Omega \text{ cm}^2$. The ohmic contact is formed in an n -type silicon. If $N_D = 5 \times 10^{19} \text{ cm}^{-3}$, $\phi_{Bn} = 0.8 \text{ V}$, and the electron effective mass is $0.26 m_0$, find the voltage drop across the contact when a forward current of 1A flows through it.

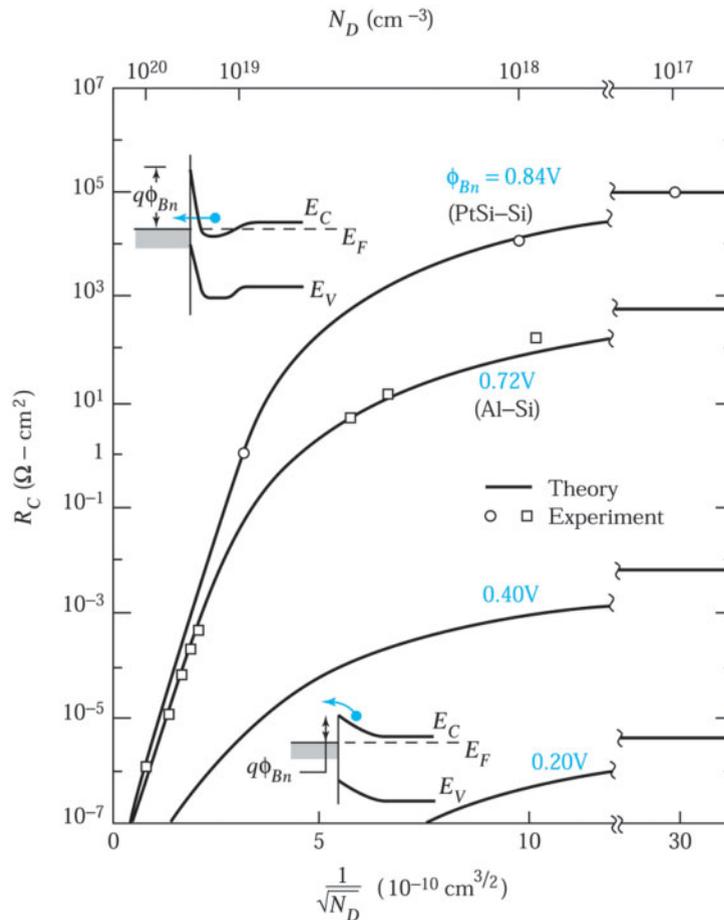


Fig. 9 Calculated and measured values of specific contact resistance. Upper inset shows the tunneling process, lower inset shows thermionic emission over the low barrier.⁶

SOLUTION The contact resistance for the ohmic contact is

$$\frac{R_c}{A} = 10^{-6} \Omega \cdot \text{cm}^2 / 10^{-5} \text{cm}^2 = 10^{-1} \Omega,$$

$$C_2 = 4\sqrt{m_r \epsilon_s} / \hbar = \frac{4\sqrt{0.26 \times 9.1 \times 10^{-31} \times (1.05 \times 10^{-10})}}{1.05 \times 10^{-34}}$$

$$= 1.9 \times 10^{14} (\text{m}^{3/2} / \text{V}).$$

From Eq. 22,

$$I = I_0 \exp\left[-\frac{C_2(\phi_{Bn} - V)}{\sqrt{N_D}}\right],$$

$$\frac{\partial I}{\partial V}\bigg|_{V=0} = \frac{A}{R_c} = I_0 \left(\frac{C_2}{\sqrt{N_D}}\right) \exp\left(\frac{-C_2\phi_{Bn}}{\sqrt{N_D}}\right)$$

or

$$I_0 = \frac{A}{R_c} \left(\frac{\sqrt{N_D}}{C_2}\right) \exp\left(\frac{C_2\phi_{Bn}}{\sqrt{N_D}}\right)$$

$$= 10 \times \left(\frac{\sqrt{5 \times 10^{19} \times 10^6}}{1.9 \times 10^{14}}\right) \exp\left(\frac{1.9 \times 10^{14} \times 0.8}{\sqrt{5 \times 10^{19} \times 10^6}}\right)$$

$$= 8.13 \times 10^8 \text{ A}.$$

At $I = 1 \text{ A}$, we have

$$\phi_{Bn} - V = \frac{\sqrt{N_D}}{C_2} \ln\left(\frac{I_0}{I}\right) = 0.763 \text{ V}$$

or

$$V = 0.8 - 0.763 = 0.037 \text{ V} = 37 \text{ mV}.$$

Therefore, there is a negligibly small voltage drop across the ohmic contact. However, the voltage drop may become significant when the contact area is reduced to 10^{-8}cm^2 or smaller. ◀

► 7.2 MESFET

7.2.1 Basic Device Structures

The metal-semiconductor field-effect transistor (MESFET) was proposed⁷ in 1966. The MESFET has three metal-semiconductor contacts—one Schottky barrier for the gate electrode and two ohmic contacts for the source and drain electrodes. A perspective view of a MESFET is illustrated in Fig. 10a. The basic device parameters include L , the gate length, Z , the gate width, and a , the thickness of the epitaxial layer. Most MESFETs are made of n -type III-V compound semiconductors, such as gallium arsenide, because of their high electron mobilities, which help to minimize series resistances, and because of their high saturation velocities, which result in the increase in the cutoff frequency.

Practical MESFETs are fabricated by using epitaxial layers on semiinsulating substrates to minimize parasitic capacitances. In Fig. 10a, the ohmic contacts are labeled source and drain, and the Schottky barrier is labeled gate. A MESFET is often described in terms of the gate dimensions. If the gate length (L) is $0.5 \mu\text{m}$ and the gate width (Z) is $300 \mu\text{m}$, the device is referred to as a $0.5 \times 300 \mu\text{m}$ device. A microwave- or millimeter-wave device typically has a gate length in the range 0.1 – $1.0 \mu\text{m}$. The thickness a of the epitaxial layer is typically one-third to one-fifth of the gate length. The spacing between the electrodes is one to four times that of the gate length. The current handling capability of a MESFET is directly proportional to the gate width Z because the cross-sectional area available for channel current is proportional to Z .

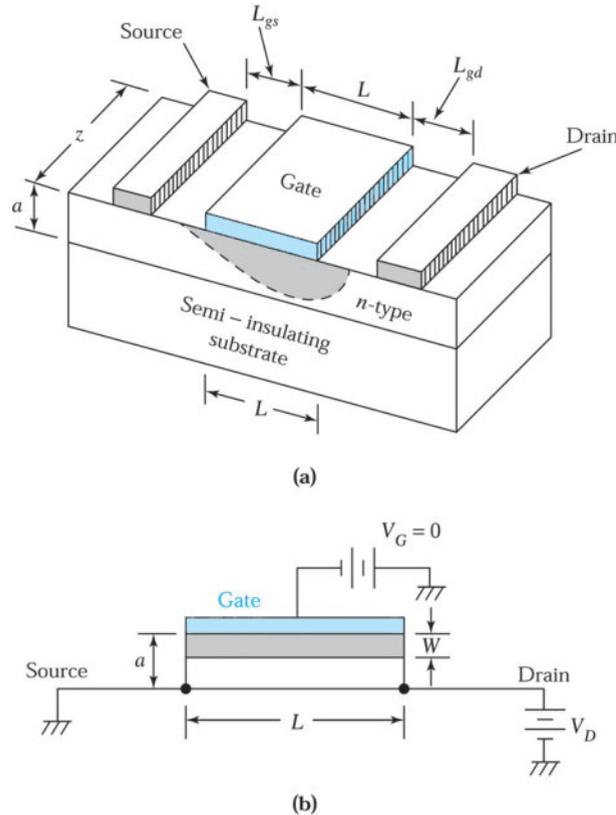


Fig. 10 (a) Perspective view of a metal-semiconductor field-effect transistor (MESFET). (b) Cross section of the gate region of a MESFET.

7.2.2 Principles of Operation

To understand the operation of a MESFET, we consider the section under the gate, Fig. 10b. The source is grounded, and the gate and drain voltage are measured with respect to the source. Under normal operating conditions, the gate voltage is zero or reverse biased and the drain voltage is zero or forward biased; that is, $V_G \leq 0$ and $V_D \geq 0$. Since the channel is n -type material, the device is referred to as an n -channel MESFET. Most applications use the n -channel MESFET rather than the p -channel MESFET because of higher carrier mobility in n -channel devices.

The resistance of the channel is given by

$$R = \rho \frac{L}{A} = \frac{L}{q\mu_n N_D A} = \frac{L}{q\mu_n N_D Z(a-W)}, \quad (24)$$

where N_D is the donor concentration, A is the cross-section area for current flow and equals $Z(a-W)$, and W is the width of the depletion region of the Schottky barrier.

When no gate voltage is applied and V_D is small, as shown in Fig. 11a, a small drain current I_D flows in the channel. The magnitude of the current is given by V_D/R , where R is the channel resistance given in Eq. 24. Therefore, the current varies linearly with the drain voltage. Of course, for any given drain voltage, the voltage along the channel increases from zero at the source to V_D at the drain. Thus, the Schottky barrier becomes increasingly reverse biased as we proceed from the source to the drain. As V_D is increased, W increases, and the average cross-sectional area for current flow is reduced. The channel resistance R also increases. As a result, the current increases at a slower rate.

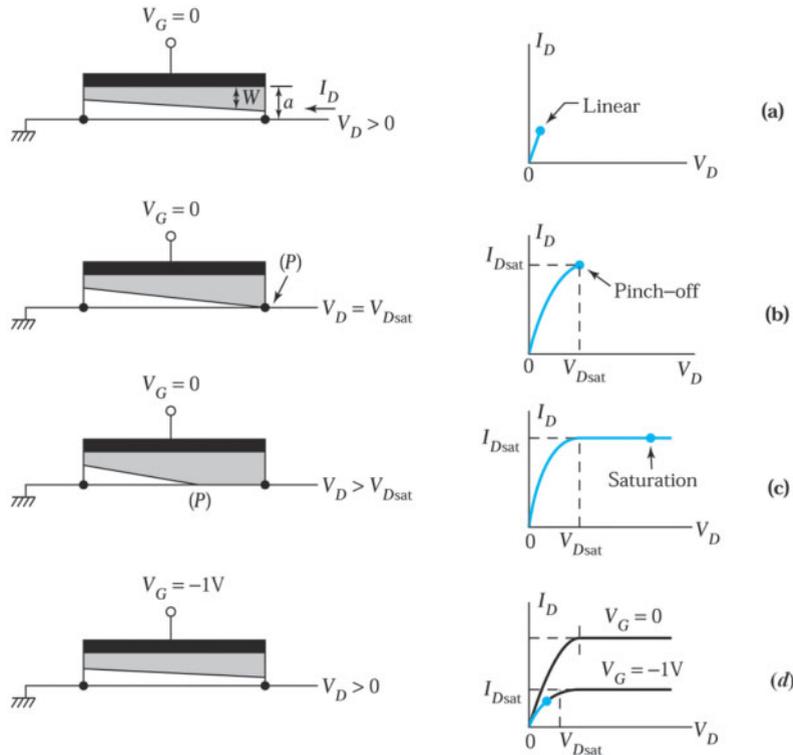


Fig. 11 Variation of the depletion-layer width and output characteristics of a MESFET under various biasing conditions. (a) $V_G = 0$ and a small V_D . (b) $V_G = 0$ and at pinch-off. (c) $V_G = 0$ at post pinch-off ($V_D > V_{Dsat}$). (d) $V_G = -1V$ and a small V_D .

As the drain voltage is further increased, eventually the depletion region touches the semiinsulating substrate, as shown in Fig. 11b. This happens when $W = a$ at the drain. We can obtain the corresponding value of the drain voltage, called the *saturation voltage*, V_{Dsat} , from Eq. 7 where $V = -V_{Dsat}$:

$$V_{Dsat} = \frac{qN_D a^2}{2\epsilon_s} - V_{bi} \quad \text{for } V_G = 0. \quad (25)$$

At this drain voltage, the source and the drain are *pinched off* or completely separated by a reverse-biased depletion region. The location P in Fig. 11b is called the pinch-off point. At this point, a large drain current called the *saturation current* I_{Dsat} can flow across the depletion region. This is similar to the situation caused by injecting carriers into a reverse-biased depletion region such as the collector-base depletion region of a bipolar transistor.

Beyond the pinch-off point, as V_D is increased further, the depletion region near the drain will expand and point P will move toward the source, as indicated in Fig. 11c. However, the voltage at point P remains the same, V_{Dsat} . Thus, the number of electrons per unit time arriving from the source to point P , and hence the current flowing in the channel, remain the same because the potential drop in the channel from source to point P does not change. Therefore, for drain voltages larger than V_{Dsat} , the current remains essentially at the value I_{Dsat} and is independent of V_D .

When a gate voltage is applied to reverse bias the gate contact, the depletion-layer width W increases. For a small V_D , the channel again acts as a resistor but its resistance is higher because the cross-sectional area available for current flow is decreased. As indicated in Fig. 11d, the initial current is smaller for $V_G = -1V$ than for $V_G = 0$. When V_D is increased to a critical value, the depletion region again touches the semiinsulating substrate. The value of this V_D is given by

$$V_{Dsat} = \frac{qN_D a^2}{2\epsilon_s} - V_{bi} - V_G. \quad (26)$$

For an n -channel MESFET, the gate voltage is negative with respect to the source, so we use the absolute value of V_G in Eq. 26 and in subsequent equations. We see from Eq. 26 that the application of a gate voltage V_G reduces the drain voltage required for the onset of pinch-off by an amount equal to V_G .

7.2.3 Current-Voltage Characteristics

We now consider a MESFET before the onset of pinch-off, as shown in Fig. 12a. The drain voltage variation along the channel is shown in Fig. 12b. The voltage drop across an elemental section dy of the channel is given by

$$dV = I_D dR = \frac{I_D dy}{q\mu_n N_D Z [a - W(y)]}, \quad (27)$$

where we used Eq. 24 for dR and replaced L by dy . The depletion-layer width at distance y from the source is given by

$$W(y) = \sqrt{\frac{2\epsilon_s [V(y) + V_G + V_{bi}]}{qN_D}}. \quad (28)$$

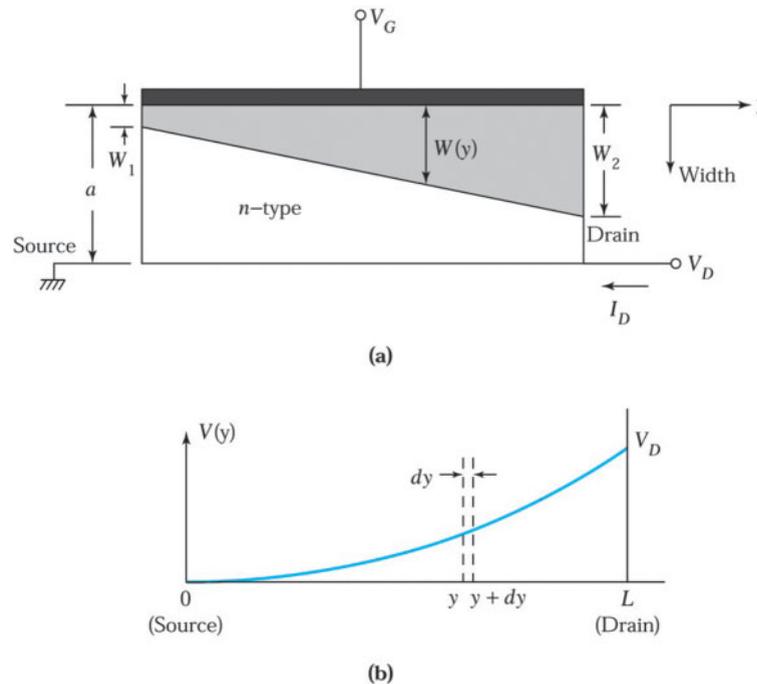


Fig. 12 (a) Expanded view of the channel region. (b) Drain voltage variation along the channel.

The drain current I_D is a constant, independent of y . We can rewrite Eq. 27 as

$$I_D dy = q\mu_n N_D Z [a - W(y)] dV. \quad (29)$$

The differentiation of the drain voltage dV is obtained from Eq. 28:

$$dV = \frac{qN_D}{\epsilon_s} W dW. \quad (30)$$

Substituting dV into Eq. 29 and integrating from $y = 0$ to $y = L$ yields

$$\begin{aligned} I_D &= \frac{1}{L} \int_{w_1}^{w_2} q\mu_n N_D Z (a - W) \frac{qN_D}{\epsilon_s} W dW \\ &= \frac{Z\mu_n q^2 N_D^2}{2\epsilon_s L} \left[a(W_2^2 - W_1^2) - \frac{2}{3}(W_2^3 - W_1^3) \right] \end{aligned}$$

or

$$I = I_P \left[\frac{V_D}{V_P} - \frac{2}{3} \left(\frac{V_D + V_G + V_{bi}}{V_P} \right)^{3/2} + \frac{2}{3} \left(\frac{V_G + V_{bi}}{V_P} \right)^{3/2} \right], \quad (31)$$

where

$$I_P \equiv \frac{Z\mu_n q^2 N_D^2 a^3}{2\epsilon_s L} \quad (31a)$$

and

$$V_P \equiv \frac{qN_D a^2}{2\epsilon_s}. \quad (31b)$$

The voltage V_P is called the pinch-off voltage, that is, the total voltage ($V_D + V_G + V_{bi}$) at which $W_2 = a$.

In Fig. 13 we show the I - V characteristics of a MESFET having a pinch-off voltage of 3.2 V. The curves shown are calculated for $0 \leq V_D \leq V_{Dsat}$ using Eq. 31. Beyond V_{Dsat} the current is taken to be constant in accordance with our previous discussion. Note that there are three different regions in the current-voltage characteristics. When V_D is small, the cross-section area of the channel is essentially independent of V_D and the I - V characteristics are ohmic or linear. We refer to this region of operation as the linear region. For $V_D \geq V_{Dsat}$, the current saturates at I_{Dsat} . We refer to this region of operation as the saturation region. As the drain voltage is further increased, avalanche breakdown of the gate-to-channel diode occurs and the drain current suddenly increases. This is the breakdown region.

In the linear region where $V_D \ll V_G + V_{bi}$, Eq. 31 can be expanded to give

$$I_D \cong \frac{I_P}{V_P} \left[1 - \sqrt{\left(\frac{V_G + V_{bi}}{V_P} \right)} \right] V_D. \quad (32)$$

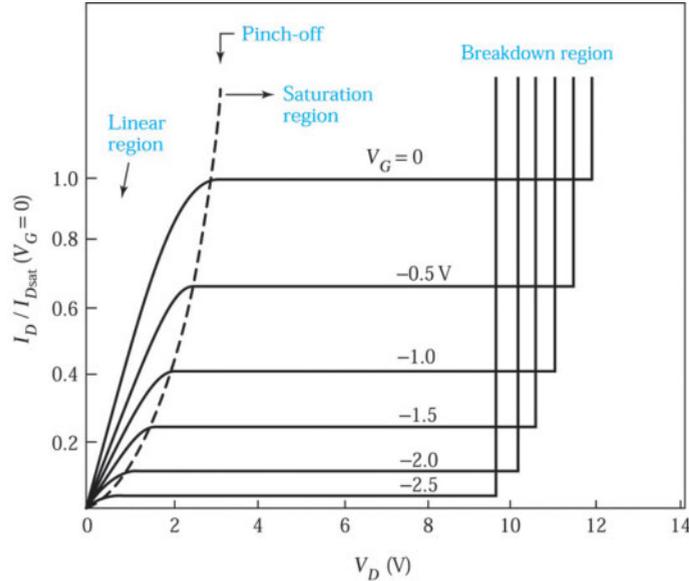


Fig. 13 Normalized ideal current-voltage characteristics of a MESFET with $V_p = 3.2$ V.

An important parameter of a MESFET is the transconductance g_m , which represents the change of drain current at a given drain voltage on a change in gate voltage. From Eq. 32, we obtain

$$g_m = \left. \frac{\partial I_D}{\partial V_G} \right|_{V_D} = \frac{I_P}{2V_P^2} \sqrt{\frac{V_P}{V_G + V_{bi}}} V_D. \quad (33)$$

In the saturation region, the drain current can be calculated from Eq. 31 by evaluating the current at the pinch-off point; that is, by setting $V_p = V_D + V_G + V_{bi}$:

$$I_{Dsat} = I_P \left[\frac{1}{3} - \left(\frac{V_G + V_{bi}}{V_P} \right) + \frac{2}{3} \left(\frac{V_G + V_{bi}}{V_P} \right)^{3/2} \right]. \quad (34)$$

The corresponding saturation voltage is given by

$$V_{Dsat} = V_P - V_G - V_{bi} \quad (35)$$

The transconductance in the saturation region can be obtained from Eq. 34:

$$g_m = \frac{I_P}{V_P} \left(1 - \sqrt{\frac{V_G + V_{bi}}{V_P}} \right) = \frac{Z\mu_n q N_D a}{L} \left(1 - \sqrt{\frac{V_G + V_{bi}}{V_P}} \right). \quad (36)$$

In the breakdown region, the breakdown voltage occurs at the drain end of the channel, where the reverse voltage is the highest:

$$V_B \text{ (breakdown voltage)} = V_D + |V_G| \quad (37)$$

For example, in Fig. 13 the breakdown voltage is 12 V for $V_G = 0$. At $|V_G| = 1$, the breakdown voltage is still 12 V and the drain voltage at breakdown is $(V_B - |V_G|)$ or 11 V.

► EXAMPLE 4

Consider an n -channel GaAs MESFET at $T = 300$ K with a gold contact. Assume the barrier height is 0.89 V. The n -channel doping is $2 \times 10^{15} \text{ cm}^{-3}$ and the channel thickness is $0.6 \mu\text{m}$. Calculate the pinch-off voltage and the built-in potential. The dielectric constant of GaAs is 12.4.

SOLUTION The pinch-off voltage is

$$V_p = \frac{qN_D}{2\epsilon_s} a^2 = \frac{(1.6 \times 10^{-19})(2 \times 10^{15})}{2 \times 12.4 \times (8.85 \times 10^{-14})} \times (0.6 \times 10^{-4})^2 = 0.53 \text{ V.}$$

The difference between the conduction band and the Fermi level is given by

$$V_n = \frac{kT}{q} \ln \left(\frac{N_C}{N_D} \right) = 0.026 \ln \left(\frac{4.7 \times 10^{17}}{2 \times 10^{15}} \right) = 0.14 \text{ V.}$$

The built-in potential is

$$V_{bi} = \phi_{Bn} - V_n = 0.89 - 0.14 = 0.75 \text{ V.} \quad \blacktriangleleft$$

So far we have considered only a normally on (or depletion-mode) device; that is, the device has a conductive channel at $V_G = 0$. For high-speed, low-power applications, the normally off device is preferred. This device does not have a conductive channel at $V_G = 0$; that is, the built-in potential V_{bi} of the gate contact is sufficient to deplete the channel region. This is possible, for example, in a gallium arsenide MESFET with a very thin epitaxial layer on a semiinsulating substrate. For a normally off MESFET, a positive bias must be applied to the gate before channel current begins to flow. The required voltage, called the *threshold voltage* V_T , is given by

$$V_T = V_{bi} - V_p \quad (38a)$$

or

$$V_{bi} = V_T + V_p, \quad (38b)$$

where V_p is the pinch-off voltage defined in Eq. 31b. Near the threshold voltage, the drain current in the saturation region can be obtained by substituting V_{bi} of Eq. 38b in Eq. 34 and by using the Taylor series expansion assuming $(V_G - V_T)/V_p \ll 1$. We obtain

$$I_{Dsat} = I_p \left\{ \frac{1}{3} - \left[1 - \left(\frac{V_G - V_T}{V_p} \right) \right] + \frac{2}{3} \left[1 - \left(\frac{V_G - V_T}{V_p} \right) \right]^{3/2} \right\}$$

or

$$I_{Dsat} \approx \frac{Z\mu_n\epsilon_s}{2aL} (V_G - V_T)^2. \quad (39)$$

In deriving Eq. 39 we used a negative sign for V_G to take into account its polarity.

The basic current-voltage characteristics of normally on and normally off devices are similar. Figure 14 compares these two modes of operation. The main difference is the shift of threshold voltage along the V_G axis. The normally off device (Fig. 14b) has no current conduction at $V_G = 0$, and the current varies as in Eq. 39 when $V_G > V_T$. Since the built-in potential of the gate is less than about 1 V, the forward bias on the gate is limited to about 0.5 V to avoid excessive gate current.

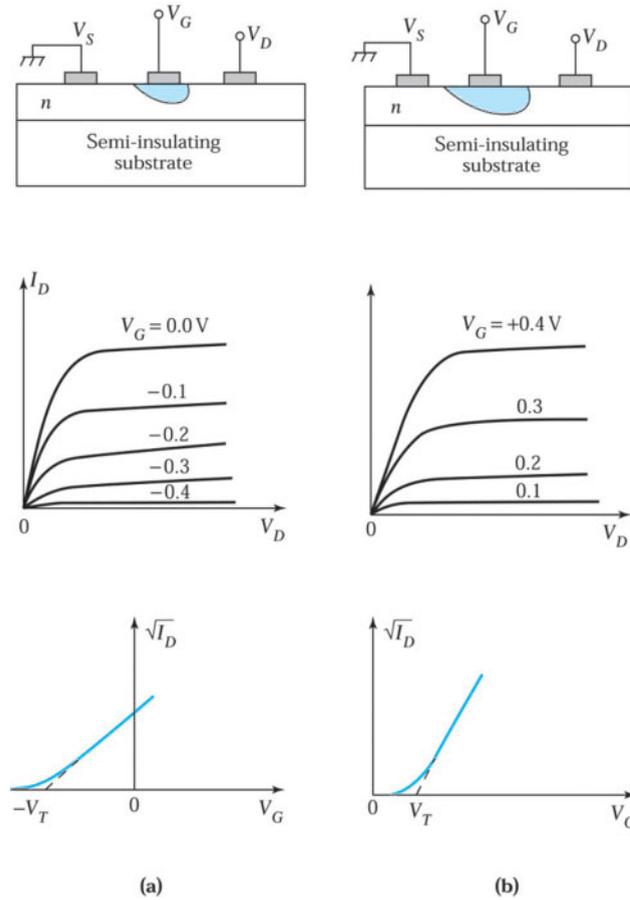


Fig. 14 Comparison of I-V characteristics. (a) Normally on MESFET. (b) Normally off MESFET.

The transconductance for a normally off device can be obtained from Eq. 39:

$$g_m = \frac{dI_{Dsat}}{dV_G} = \frac{Z\mu_n\epsilon_s}{aL}(V_G - V_T). \quad (40)$$

7.2.4 High-Frequency Performance

For high-frequency applications of MESFETs, an important figure of merit is the cutoff frequency f_T , which is the frequency at which the MESFET can no longer amplify the input signal. The small-signal input current is the product of the gate admittance and the small-signal gate voltage, assuming that the device has negligibly small series resistance:

$$\tilde{i}_{in} = 2\pi f C_G \tilde{v}_g, \quad (41)$$

where C_G is the gate capacitance equal to $ZL(\epsilon_s/\bar{W})$ and \bar{W} is the average depletion-layer width under the gate electrode. The small-signal output current is obtainable from the definition of the transconductance:

$$g_m = \frac{\partial I_D}{\partial V_G} = \frac{\tilde{i}_{\text{out}}}{\tilde{v}_g} \quad (42)$$

or

$$\tilde{i}_{\text{out}} = g_m \tilde{v}_g.$$

Equating Eqs. 41 and 42a, we obtain the cutoff frequency

$$f_T = \frac{g_m}{2\pi C_G} < \frac{I_P / V_P}{2\pi ZL(\epsilon_s / W)} \approx \frac{\mu_n q N_D a^2}{2\pi \epsilon_s L^2}, \quad (43)$$

where we used Eq. 36 for g_m . From Eq. 43 we see that to improve high-frequency performance, we should use a MESFET having high carrier mobility and short channel length. This is the reason that the n -channel MESFET, which has higher electron mobility, is preferred.

These derivations are based on the assumption that the carrier mobility in the channel is a constant independent of the applied field. However, for very-high-frequency operations, the longitudinal field, i.e., the electric field directed from the source to the drain, is sufficiently high that the carriers travel at their saturation velocity.

Under these conditions, the saturation channel current is given by

$$\begin{aligned} I_{D\text{sat}} &= (\text{area for carrier transport}) \times qm v_s, \\ &= Z(a - W)qN_D v_s. \end{aligned} \quad (44)$$

The transconductance is then

$$g_m = \frac{\partial I_{D\text{sat}}}{\partial V_G} = \frac{\partial I_{D\text{sat}}}{\partial W} \cdot \frac{\partial W}{\partial V_G} = [qN_D v_s Z(-1)] \left(\frac{1}{-qN_D W / \epsilon_s} \right) \quad (45)$$

or

$$g_m = Z v_s \epsilon_s / W. \quad (45a)$$

In Eq. 45, we obtain $\partial W / \partial V_G$ from Eq. 28.

From Eq. 45a, we can obtain the cutoff frequency under saturation-velocity condition:

$$\boxed{f_T = \frac{g_m}{2\pi C_G} = \frac{Z v_s \epsilon_s / W}{2\pi ZL(\epsilon_s / W)} = \frac{v_s}{2\pi L}.} \quad (46)$$

Therefore, to increase f_T , we must reduce the gate length L and employ a semiconductor with a high saturation velocity. Figure 15 shows the electron drift velocity versus electric field for five semiconductors.⁸ Note that GaAs has an average velocity[§] of 1.2×10^7 cm/s and a peak velocity of 2×10^7 cm/s, which are 20%–100% higher than the saturation velocity of silicon. Also note that Ga_{0.47}In_{0.53}As and InP have higher average and peak velocities than GaAs. Consequently, the cutoff frequencies of these semiconductors will be higher than that from GaAs.

[§] The average velocity is defined as $\bar{v} \equiv \left[\frac{1}{L} \int_0^L \frac{dx}{v(x)} \right]^{-1}$. If $v(x)$ is a constant v_0 , then $\bar{v} = v_0$.

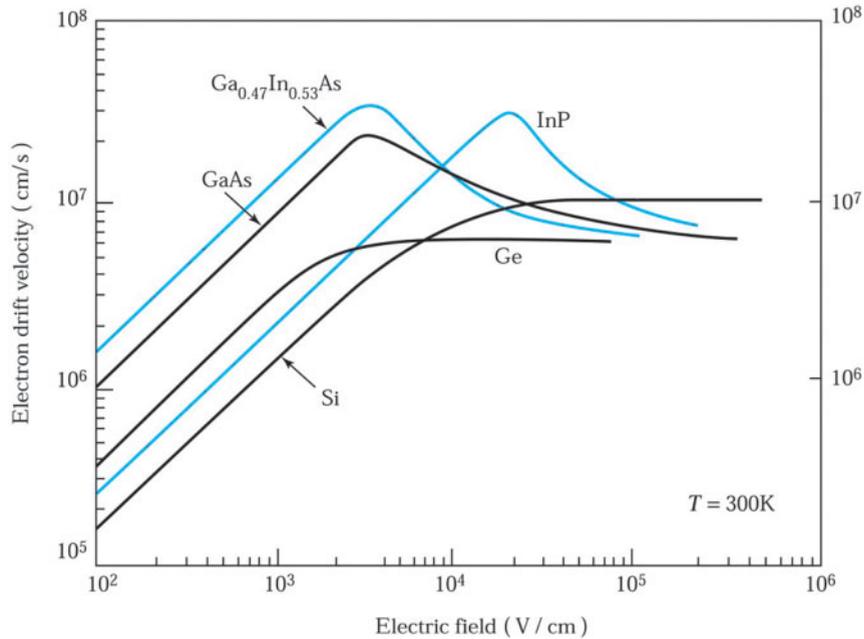


Fig. 15 The drift velocity versus the electric field for electrons in various semiconductor materials.⁸

► 7.3 MODFET

7.3.1 MODFET Fundamentals

The modulation-doped field-effect transistor (MODFET) is a heterostructure field-effect device. The MODFET has been an alternative to MESFETs in high-speed circuits. Other names commonly applied to the device include high electron mobility transistor (HEMT), two-dimensional electron gas field-effect transistor (TEGFET), and selectively doped heterostructure transistor (SDHT). Frequently, it is referred to by a general name of heterojunction field-effect transistor (HFET).

The most-common heterojunctions for the MODFETs are the AlGaAs/GaAs, AlGaAs/InGaAs, and InAlAs/InGaAs heterointerfaces. Figure 16 shows a perspective view of a conventional AlGaAs/GaAs MODFET. The special features of a MODFET are its heterojunction structure under the gate and its modulation doped layers. For the device in Fig. 16, AlGaAs is the wide bandgap semiconductor, whereas GaAs is the narrow bandgap semiconductor. The two semiconductors are modulation doped, i.e., the AlGaAs is doped ($\sim 10^{18} \text{ cm}^{-3}$), except for a narrow region d_s , which is undoped, whereas the GaAs is undoped. Electrons in the AlGaAs will diffuse to the undoped GaAs, where a conduction channel can be formed at the surface of the GaAs. The net result of this modulation doping is that channel carriers have high mobilities because there is no impurity scattering. The undoped AlGaAs spacer layer is used to reduce the Coulomb scattering from ionized donors in doped AlGaAs and leads to enhanced carrier mobility in the channel.

A comparison of low-field electron mobility of the modulation-doped 2-D channel to bulk GaAs at different doping levels is shown in Fig. 17. In a MESFET the channel has to be doped to a reasonably high level ($> 10^{17} \text{ cm}^{-3}$), and thus the electron suffers from impurity scattering. However, the modulation-doped channel has much higher mobilities at all temperatures. It is also interesting to compare the modulation-doped channel, which usually has an unintentional doping below 10^{14} cm^{-3} , to lowly doped bulk samples (with a similar impurity concentration of $4 \times 10^{13} \text{ cm}^{-3}$). The bulk mobility as a function of temperature shows a peak but drops at both

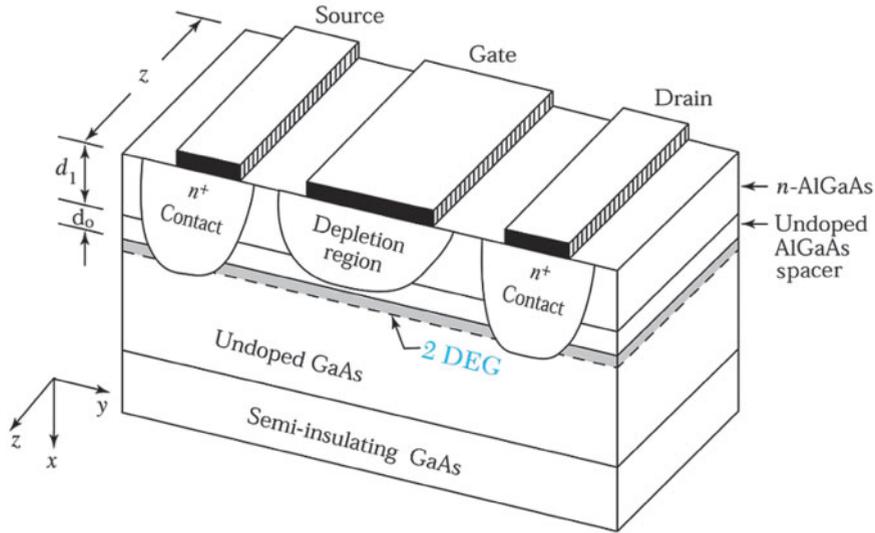


Fig. 16 Perspective view of a conventional modulation-doped field-effect transistor (MODFET) structure.

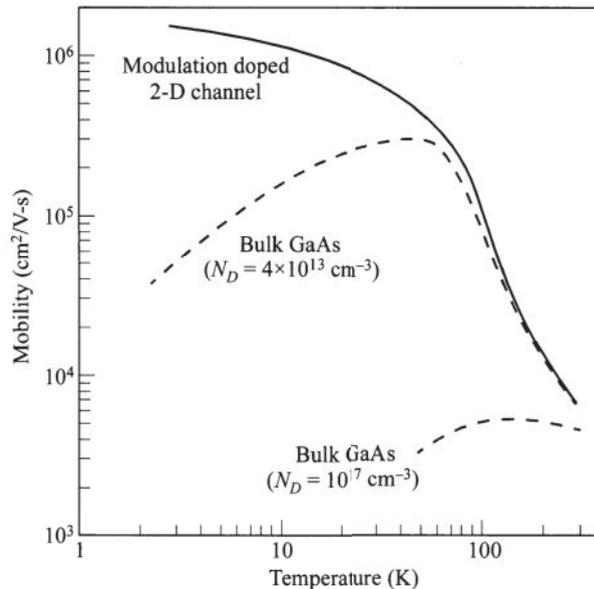


Fig.17 Comparison of low-field electron mobility of modulation-doped 2-D channel to bulk GaAs at different doping levels.⁹

high temperature and low temperature. The decrease of bulk mobility with an increase of temperature is due to phonon scattering. At low temperatures, the bulk mobility is limited by impurity scattering. It depends on the doping level and also decreases with a decrease in temperature. In the modulation-doped channel, its mobility at temperatures above ~ 80 K is comparable to the value of a lowly doped bulk sample. However, mobility is much enhanced at lower temperatures. The modulation-doped channel does not suffer from impurity scattering,

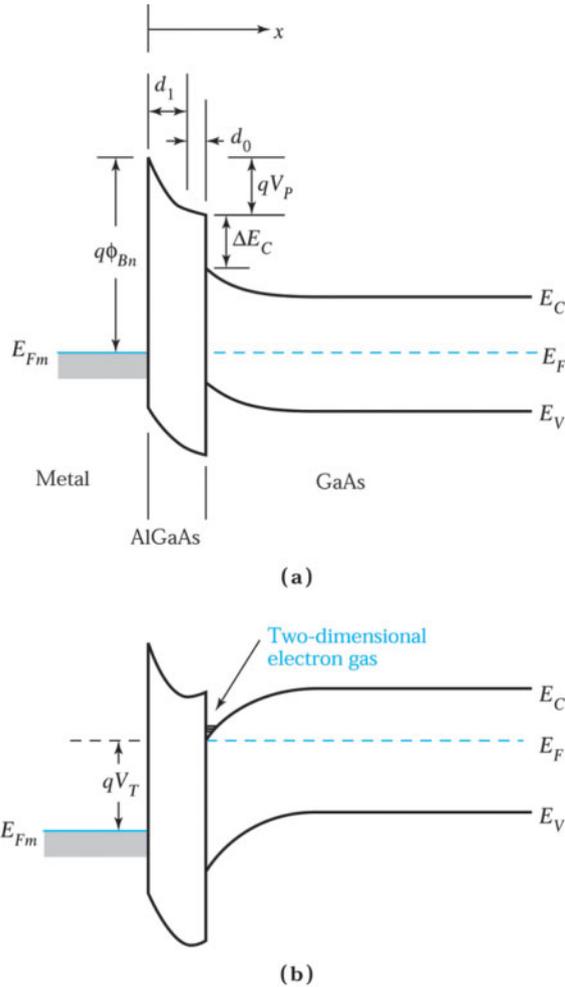


Fig. 18 Energy band diagrams for a normally-off MODFET at (a) thermal equilibrium, and (b) the onset of threshold. d_1 and d_0 are the doped and undoped regions, respectively.⁶

which dominates at low temperatures. This benefit stems from the screening effect of a *two-dimensional electron gas* (2 DEG), where its conduction path is confined to a small cross-section, smaller than 10 nm, with high volume density.

Figure 18a shows the band diagram of a MODFET in a thermal equilibrium condition. Similar to a standard Schottky barrier, $q\phi_{Bn}$ is the barrier height of the metal on the wide-bandgap semiconductor.¹⁰ ΔE_C is the conduction band discontinuity for the heterojunction structure, and V_p is the built-in potential given by

$$V_p = \frac{q}{\epsilon_s} \int_0^{d_1} N_D(x) x dx = \frac{qN_D d_1^2}{2\epsilon_s}, \quad (47)$$

where d_1 is the thickness of the doped region in AlGaAs and ϵ_s is the dielectric permittivity.

A key parameter for the operation of a MODFET is the threshold voltage V_T , which is the gate bias at which the channel starts to form between the source and drain. With reference to Fig. 18b, V_T corresponds to the situation where the bottom of the conduction band at the GaAs surface coincides with the Fermi level:

$$V_T = \phi_{Bn} - \frac{\Delta E_C}{q} - V_p. \quad (48)$$

The threshold voltage V_T can be adjusted by using different values for ϕ_{Bn} and V_p . However, ΔE_C is fixed for a given set of semiconductors. Figure 18b has a positive V_T , and the MODFET is an enhancement-mode device (normally off), as opposed to a depletion-mode device with a negative V_T (normally on).

When the gate voltage is larger than V_T , a charge sheet $n_s(y)$ is capacitively induced by the gate at the heterojunction interface. The charge sheet is similar to the charge Q_n/q in the inversion layer of a MOSFET (see Section 5.1):

$$n_s(y) = \frac{C_i [V_G - V_T - V(y)]}{q}, \quad (49)$$

where

$$C_i = \frac{\epsilon_s}{d_1 + d_0 + \Delta d}, \quad (49a)$$

d_1 and d_0 are the doped and undoped AlGaAs thickness (Fig. 16) and Δd is the channel thickness or the thickness of the inversion layer, estimated to be about 8 nm. $V(y)$ is the channel potential with respect to the source. The channel potential varies along the channel from 0 to the drain bias V_D , similar to that shown in Fig. 12b. The charge sheet is also called a two-dimensional electron gas. This is because that the electrons in the inversion layer are confined in the x -direction by ΔE_C on the left side and by the potential distribution of the conduction band on the right side (Fig. 18b). However, these electrons can make two-dimensional movements: in the y -direction from the source to the drain and in the z -direction parallel to the channel width (Fig. 16).

Equation 49 shows that a negative gate bias will reduce the two-dimensional electron gas. If, on the other hand, a positive V_G is applied, n_s will increase.

► EXAMPLE 5

Consider an AlGaAs/GaAs heterojunction with n -AlGaAs doped to $2 \times 10^{18} \text{ cm}^{-3}$ and a thickness of 40 nm. Assume the undoped spacer layer is 3 nm and the Schottky barrier height is 0.85 V and $\frac{\Delta E_C}{q} = 0.23 \text{ V}$. The dielectric constant of the AlGaAs is 12.3. Calculate the two-dimensional electron gas concentration for such heterojunction at $V_G = 0$.

SOLUTION

$$V_p = \frac{qN_D d_1^2}{2\epsilon_s} = \frac{1.6 \times 10^{-19} \times 2 \times 10^{18} \times (40 \times 10^{-7})^2}{2 \times 12.3 \times 8.85 \times 10^{-14}} = 2.35 \text{ V}.$$

The threshold voltage is

$$V_T = \phi_{Bn} - \frac{\Delta E_C}{q} - V_p = 0.85 - 0.23 - 2.35 = -1.73 \text{ V}.$$

Therefore, the device is a normally on MODFET.

The two-dimensional electron gas at the source for $V_G = 0$ is

$$n_s = \frac{12.3 \times 8.85 \times 10^{-14}}{1.6 \times 10^{-19} \times (40 + 3 + 8) \times 10^{-7}} \times [0 - (-1.73)] = 2.29 \times 10^{12} \text{ cm}^{-2}. \quad \blacktriangleleft$$

7.3.2 Current-Voltage Characteristics

The current-voltage characteristics of a MODFET can be obtained by using the gradual channel approximation similar to that of a MOSFET. The current at any point along the channel is given by

$$\begin{aligned} I &= Zq\mu_n n_s \mathcal{E}_y \\ &= Z\mu_n C_i [V_G - V_T - V(y)] \frac{dV(y)}{dy}. \end{aligned} \quad (50)$$

Since the current is constant along the channel, integrating Eq. 50 from source to drain ($y = 0$ to $y = L$) gives

$$I = \frac{Z}{L} \mu_n C_i \left[(V_G - V_T) V_D - \frac{V_D^2}{2} \right]. \quad (51)$$

The output characteristics for an enhancement-mode MODFET are similar to those shown in Fig. 14b. In the linear region where $V_D \ll (V_G - V_T)$, Eq. 51 can be reduced to

$$I = \frac{Z}{L} \mu_n C_i (V_G - V_T) V_D. \quad (52)$$

For a large drain voltage, the charge sheet $n(y)$ at the drain is reduced to zero. This is the pinch-off condition previously discussed and shown in Fig. 11b. From Eq. 49, we obtain the saturation voltage V_{Dsat} , at which $n_s(y = L) = 0$,

$$V_{Dsat} = V_G - V_T, \quad (53)$$

and the saturation current can be obtained from Eqs. 51 and 53,

$$I = \frac{Z\mu_n C_i}{2L} (V_G - V_T)^2 = \frac{Z\mu_n \epsilon_s}{2L(d_1 + d_0 + \Delta d)} (V_G - V_T)^2. \quad (54)$$

Note that this equation is very similar to Eq. 39. A similar expression can be obtained for the transconductance given in Eq. 40.

For high-speed operations, the longitudinal field (i.e., the field along the channel) is sufficiently high to cause carrier velocity saturation. The current in the velocity-saturation region is

$$\begin{aligned} I_{sat} &= Zv_s qn_s \\ &\cong Zv_s C_i (V_G - V_T). \end{aligned} \quad (55)$$

The transconductance becomes

$$g_m = \frac{\partial I_{sat}}{\partial V_G} = Zv_s C_i. \quad (56)$$

Note that I_{sat} is independent of gate length and g_m is independent of both gate length and gate voltage in the velocity-saturation regime.

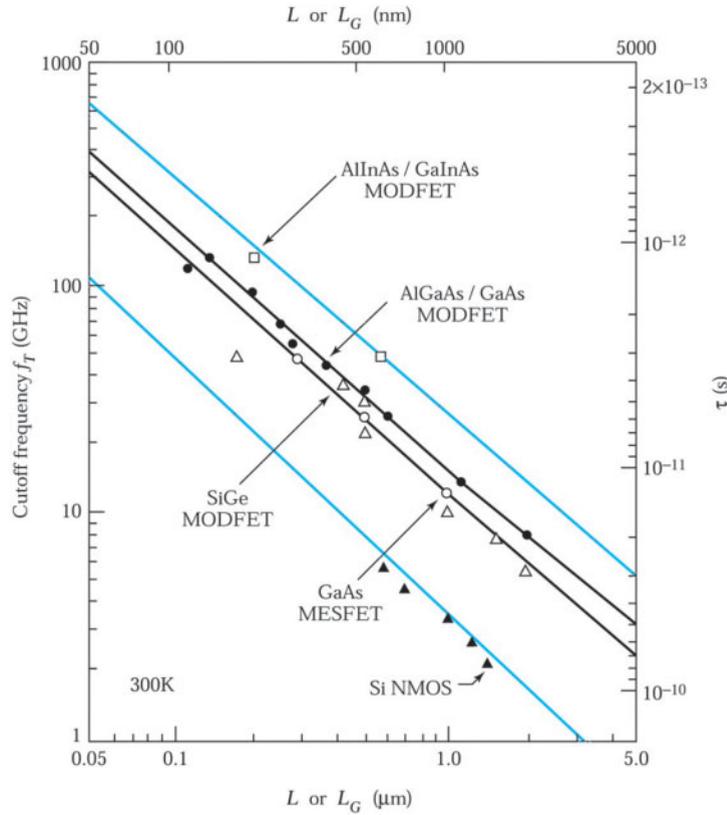


Fig. 19 Cutoff frequency versus channel or gate length for five different field-effect transistors.^{8,11}

7.3.3 Cutoff Frequency

The speed of a MODFET is measured by the cutoff frequency

$$\begin{aligned}
 f_T &= \frac{g_m}{2\pi(\text{total capacitance})} = \frac{Zv_s C_i}{2\pi(ZLC_i + C_p)} \\
 &= \frac{v_s}{2\pi(L + C_p / ZC_i)}.
 \end{aligned} \tag{57}$$

where C_p is the parasitic capacitance. To improve f_T , we should consider a semiconductor with a large v_s , a gate structure with an ultrashort gate length, and a device configuration with minimum parasitic capacitance.

A comparison of the cutoff frequencies of various FETs is shown in Fig. 19. The cutoff frequency f_T is plotted against the channel or gate length.^{8,11} Note that for a given length, a silicon n -type MOSFET has the lowest f_T because of the relatively low mobility and a low average velocity of electron in silicon. GaAs MESFET has an f_T about three times higher than does a silicon MOSFET.

Also shown are three MODFETs. The conventional GaAs MODFET (i.e., AlGaAs-GaAs structure) has an f_T about 30% higher than that of the GaAs MESFET. The pseudomorphic SiGe MODFET (i.e., Si-SiGe structure where the SiGe lattice is shrunk slightly to match the silicon lattice) has an f_T comparable to the GaAs MODFET. SiGe MODFETs are attractive because they can be processed in a silicon fabrication facility. For even higher cutoff frequencies, we have the $\text{Al}_{0.48}\text{In}_{0.52}\text{As-Ga}_{0.47}\text{In}_{0.53}\text{As}$ MODFET formed on an InP substrate. The superior performance is mainly due to the high electron mobility in $\text{Ga}_{0.47}\text{In}_{0.53}\text{As}$ and its high average and peak velocities. It is estimated that at a gate length of 50 nm f_T can be as high as 600 GHz.

► SUMMARY

When a metal makes intimate contact with a semiconductor, it forms a *metal-semiconductor contact*. There are two types of contacts. The first type is the rectifying contact, also called the Schottky barrier contact, which has a relatively large barrier height and is formed on a semiconductor with a relatively low doping concentration. The potential and field distribution in a Schottky barrier are identical to those of a one-sided abrupt $p-n$ junction. However, the current transport in a Schottky barrier is by thermionic emission and, therefore, has an inherent fast response.

The second type is the ohmic contact, which is formed on a degenerate semiconductor in which carrier transport is by the tunneling process. An ohmic contact can pass the required current with a very small voltage drop across it. All semiconductor devices and integrated circuits need ohmic contacts to make connections to other devices in an electronic system.

Metal-semiconductor contacts are building blocks for MESFET and MODFET devices. By employing a Schottky barrier as the gate electrode and two ohmic contacts as the source and drain electrodes, we form a MESFET. This three-terminal device is important for high-frequency applications, especially for monolithic microwave integrated circuits (MMIC). Most MESFETs are made with n -type III-V compound semiconductors because of their high electron mobilities and high average drift velocities. Of particular importance is GaAs, because of its relatively mature technology and the availability of high-quality GaAs wafers.

The MODFET is a device with enhanced high-frequency performance. This device structure is similar to that of a MESFET except there is a heterojunction under the gate. A two-dimensional electron gas, i.e., a conductive channel, is formed at the heterojunction interface, and electrons with high mobility and high average drift velocity can be transported from the source through the channel to the drain.

The output characteristics of all field-effect transistors (FETs) are similar. They all have a linear region at low-drain biases. As the bias increases, the output current eventually saturates, and at a sufficiently high voltage, avalanche breakdown occurs at the drain. Depending on whether it requires a positive- or negative-threshold voltage, FET can be either normally off (enhancement mode) or normally on (depletion mode).

The cutoff frequency f_T is a figure of merit for the high-frequency performance of an FET. For a given length, the silicon MOSFET (n -type) has the lowest f_T and the GaAs MESFET has an f_T about three times higher than that of silicon. The conventional GaAs MODFET and the pseudomorphic SiGe MODFET have f_T about 30% higher than that of the GaAs MESFET. For even higher cutoff frequencies, we have the GaInAs MODFET, which has a projected f_T of 600 GHz at a gate length of 50 nm.

► REFERENCES

1. W. Schottky, "Halbleitertheorie der Sperrschicht," *Naturwissenschaften*, **26**, 843 (1938).
2. A. M. Cowley and S. M. Sze, "Surface States and Barrier Height of Metal Semiconductor System," *J. Appl. Phys.*, **36**, 3212 (1965).
3. G. Myburg et al., "Summary of Schottky Barrier Height Data on Epitaxially Grown n - and p -GaAs," *Thin Solid Films*, **325**, 181 (1998).
4. C. R. Crowell, J. C. Sarace, and S. M. Sze, "Tungsten-Semiconductor Schottky-Barrier Diodes," *Trans. Met. Soc. AIME*, **23**, 478 (1965).
5. V. L. Rideout, "A Review of the Theory, Technology and Applications of Metal-Semiconductor Rectifiers," *Thin Solid Films*, **48**, 261 (1978).
6. S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, 3rd Ed., Wiley Interscience, Hoboken, 2007.
7. C. A. Mead, "Schottky Barrier Gate Field-Effect Transistor," *Proc. IEEE*, **54**, 307 (1966).

8. S. M. Sze, Ed., *High Speed Semiconductor Devices*, Wiley, New York, 1992.
9. P. H. Ladbrooke, "GaAs MESFETs and High Mobility Transistors (HEMT)," in H. Thomas, D. V. Morgan, B. Thomas, J. E. Aubrey, and G. B. Morgan, Eds., *Gallium Arsenide for Devices and Integrated Circuits*, Peregrinus, London, 1986.
10. K. K. Ng, *Complete Guide to Semiconductor Devices*, McGraw Hill, New York, 1995.
11. S. Luryi, J. Xu, and A. Zaslavsky, Eds., *Future Trends in Microelectronics*, Wiley, New York, 1999.

► **PROBLEMS (* DENOTES DIFFICULT PROBLEMS)**

FOR SECTION 7.1 METAL-SEMICONDUCTOR CONTACTS

1. Calculate the theoretical barrier height and built-in potential in a metal-semiconductor diode for zero applied bias. Assume the metal work function is 4.55 eV, the electron affinity is 4.01 eV, and $N_D = 2 \times 10^{16} \text{ cm}^{-3}$ at 300 K.
2. (a) Find the donor concentration and barrier height of the W-GaAs Schottky barrier diode shown in Fig. 6. (b) Compare the barrier height with that obtained from the saturation current density of $5 \times 10^{-7} \text{ A/cm}^2$ shown in Fig. 8. (c) For a reverse bias of -1 V , calculate the depletion-layer width W , the maximum field, and the capacitance.
3. Copper is deposited on a carefully prepared n -type silicon substrate to form an ideal Schottky diode. $\phi_m = 4.65 \text{ eV}$, the electron affinity is 4.01 eV, $N_D = 3 \times 10^{16} \text{ cm}^{-3}$, and $T = 300 \text{ K}$. Calculate the barrier height, the built-in potential, the depletion-layer width, and the maximum field at a zero bias.
- *4. The capacitance of a Au- n -type GaAs Schottky barrier diode is given by the relation $1/C^2 = 1.57 \times 10^5 - 2.12 \times 10^5 V_a$, where C is expressed in μF and V_a is in volts. Taking the diode area to be 10^{-1} cm^2 , calculate the built-in potential, the barrier height, the dopant concentration, and the work function.
5. Calculate the value of V_{bi} and ϕ_m in an ideal metal-Si Schottky barrier contact. Assume the barrier height is 0.8 eV, $N_D = 1.5 \times 10^{16} \text{ cm}^{-3}$, and $q\chi = 4.01 \text{ eV}$.
6. In a metal-Si Schottky barrier contact, the barrier height is 0.75 eV and $A^* = 110 \text{ A/cm}^2 - \text{K}^2$. Calculate the ratio of the injected hole current to the electron current at 300 K, assuming $D_p = 12 \text{ cm}^2 \text{ s}^{-1}$, $L_p = 1 \times 10^{-3} \text{ cm}$, and $N_D = 1.5 \times 10^{16} \text{ cm}^{-3}$.

FOR SECTION 7.2 MESFET

7. Given $\phi_{Bn} = 0.9 \text{ eV}$ and $N_D = 10^{17} \text{ cm}^{-3}$, find the minimum value of the thickness of the epitaxial layer for a GaAs MESFET to be a depletion mode device (i.e., $V_T < 0$).
8. Assume the doping in a GaAs MESFET is $N_D = 7 \times 10^{16} \text{ cm}^{-3}$ and the dimensions are $a = 0.3 \mu\text{m}$, $L = 1.5 \mu\text{m}$, $Z = 5 \mu\text{m}$, $\mu_n = 4500 \text{ cm}^2/\text{V-s}$, and $\phi_{Bn} = 0.89 \text{ V}$. Calculate the ideal value of g_m for $V_G = 0$, and $V_D = 1 \text{ V}$.
9. The n -channel GaAs MESFET shown in Fig. 10 has a barrier height $\phi_{Bn} = 0.9 \text{ V}$, $N_D = 10^{17} \text{ cm}^{-3}$, $a = 0.2 \mu\text{m}$, $L = 1 \mu\text{m}$, and $Z = 10 \mu\text{m}$. (a) Is this an enhancement or depletion mode device? (b) Find the threshold voltage (the enhancement mode indicates $V_T > 0$; the depletion mode indicates $V_T < 0$).
10. An n -channel GaAs MESFET has a channel doping $N_D = 2 \times 10^{15} \text{ cm}^{-3}$, $\phi_{Bn} = 0.8 \text{ V}$, $a = 0.5 \mu\text{m}$, $L = 1 \mu\text{m}$, $\mu_n = 4500 \text{ cm}^2/\text{V-s}$, and $Z = 50 \mu\text{m}$. Find the pinch-off potential, threshold voltage, and the saturation current at $V_G = 0$.

11. The barrier height ϕ_{bn} of two GaAs n -channel MESFETs are the same and equal to 0.85 V. The channel doping in device 1 is $N_D = 4.7 \times 10^{16} \text{ cm}^{-3}$, and that in device 2 is $N_D = 4.7 \times 10^{17} \text{ cm}^{-3}$. Determine the channel thickness required in each device so that the threshold voltage is zero for each device.

FOR SECTION 7.3 MODFET

12. For an abrupt AlGaAs/GaAs heterojunction with the n -AlGaAs layer doped to $3 \times 10^{18} \text{ cm}^{-3}$, the Schottky barrier is 0.89 V and the heterojunction conduction-band edge discontinuity ΔE_C is 0.23 eV. Calculate the thickness of the doped AlGaAs layer d_l so that the threshold voltage is -0.5 V. Assume the permittivity of the AlGaAs is 12.3.
- *13. Find the thickness of the undoped spacer layer d_0 so that the two-dimensional electron gas concentration of an AlGaAs/GaAs heterojunction is $1.25 \times 10^{12} \text{ cm}^{-2}$ at zero gate bias. Assume that the n -AlGaAs is doped to $1 \times 10^{18} \text{ cm}^{-3}$ and has a thickness d_l of 50 nm, the Schottky barrier height is 0.89 V, and $\Delta E_C/q = 0.23$ V. The permittivity of the AlGaAs is 12.3.
14. Consider an AlGaAs/GaAs HFET with a 50 nm n -AlGaAs and a 10 nm undoped AlGaAs spacer. Assume the threshold voltage is -1.3 V, N_D is $5 \times 10^{17} \text{ cm}^{-3}$, $\Delta E_C = 0.25$ eV, the channel width is 8 nm, and the permittivity of AlGaAs is 12.3. Calculate the Schottky barrier height and the two-dimensional electron gas concentration at $V_G = 0$.
15. The AlGaAs/GaAs has a two-dimensional electron gas concentration of $1 \times 10^{12} \text{ cm}^{-2}$; the spacer is 5 nm, the channel width is 8 nm, the pinch-off voltage is 1.5 V, $\Delta E_C/q = 0.23$ V, and the doping concentration of AlGaAs is 10^{18} cm^{-3} . The Schottky barrier height is 0.8 V. Find the thickness of the doped AlGaAs and the threshold voltage.
16. Consider an n -AlGaAs–intrinsic GaAs abrupt heterojunction. Assume that the AlGaAs is doped to $N_D = 3 \times 10^{18} \text{ cm}^{-3}$ and has a thickness of 35 nm (there is no spacer). Let $\phi_{bn} = 0.89$ V, and assume that $\Delta E_C = 0.24$ eV and the dielectric constant is 12.3. Calculate (a) V_p and (b) n_s for $V_G = 0$.

Microwave Diodes; Quantum-Effect and Hot-Electron Devices

- ▶ 8.1 MICROWAVE FREQUENCY BANDS
 - ▶ 8.2 TUNNEL DIODE
 - ▶ 8.3 IMPATT DIODE
 - ▶ 8.4 TRANSFERRED-ELECTRON DEVICES
 - ▶ 8.5 QUANTUM-EFFECT DEVICES
 - ▶ 8.6 HOT-ELECTRON DEVICES
 - ▶ SUMMARY
-

Many semiconductor devices discussed in the previous chapters can be operated in the microwave region (0.1 ~ 3000 GHz). However, two-terminal devices generate the highest power levels per device area in system applications, especially at higher frequencies. Additionally, pulsed operation of these devices overcomes thermal limits and increases peak rf (radio frequency) power levels by more than an order of magnitude.¹ In this chapter, we consider some special two-terminal microwave devices including the tunnel diode, the impact ionization avalanche transit-time (IMPATT) diode, the transferred-electron device, and the resonant tunneling diode.

In the past two decades, we have witnessed considerable research and effort in the development of device structures that could exploit quantum-effect and hot-electron phenomena to enhance circuit performances. Speed is often cited as a primary benefit that quantum-effect devices (QEDs) and hot-electron devices (HEDs) can offer. The tunneling process, on which most QEDs rely, is an intrinsically fast process. In HEDs, carriers under ballistic transport can move at velocities considerably in excess of their equilibrium thermal velocity. However, a more significant advantage of QEDs and HEDs is their greater functionality. These devices can perform relatively complex circuit functions with a greatly reduced device count, replacing large numbers of transistors or passive circuit components.¹ The basic device structures and operating principles of QEDs and HEDs are discussed in this chapter.

Specifically, we cover the following topics:

- Advantages of millimeter-wave devices over those operated at lower frequencies.
- The quantum tunneling phenomenon and its related devices—tunnel diode, resonant tunneling diode (RTD), and unipolar resonant tunneling transistor.
- The IMPATT diode—the most powerful semiconductor source of millimeter-wave power.
- The transferred-electron device and its transit-time domain mode.
- The real-space-transfer transistor and its advantages as a functional device.

► 8.1 MICROWAVE FREQUENCY BANDS

The microwave frequencies cover the range from about 0.1 GHz (10^8 Hz) to 3000 GHz with corresponding wavelengths from 300 cm to 0.01 cm. For frequencies from 30 to 300 GHz, we have the millimeter-wave band because the wavelength is between 10 and 1 mm. For even higher frequencies, we have the submillimeter-wave band. The microwave frequency range is usually grouped into different bands.² The bands and corresponding frequency ranges as designated by the Institute of Electrical and Electronics Engineers (IEEE) are listed in Table 1. It is recommended that both the band and the corresponding frequency range be used when referring to microwave devices.

The development of microwave technology was driven by the demands of short-wave-length radio (and later radar) systems. The history of microwaves started with the first experiments of Heinrich Hertz around 1887. Hertz used a spark transmitter that produced signals in a very broad frequency band and he selected from these a frequency around 420 MHz with a half-wavelength antenna for his experiments. The rapid development of wireless communication products has led to an explosion in microwave technology. Since the introduction of cellular telephone service in the 1980s, there has been rapid growth of these systems as well as mobile paging devices and a variety of wireless data-communication services under the broad heading of personal communication services (PCS). In addition to these terrestrial communication systems, the field of satellite-based video, telephone, and data communication systems has also grown rapidly. These systems use the microwave frequencies from several hundred MHz to well over 60 GHz—the millimeter-wave region.³

Millimeter-wave technology offers many advantages for communications and radar systems, such as radio astronomy, clear-air turbulence detection, nuclear spectroscopy, air-traffic-control beacons, and weather radar. The advantages of millimeter waves over lower microwave and infrared systems include light weight, small size, broad bandwidths (several GHz), operation in adverse weather conditions, and narrow beamwidths with high resolution. The principal frequencies of interest in the millimeter-wave band are centered around 35, 60, 94, 140, and 220 GHz.⁴ The reason for choosing these specific frequencies is mainly the atmospheric absorption of horizontally propagated millimeter waves. The atmospheric “windows” where absorption is at a local minimum are found at about 35, 94, 140, and 220 GHz. The absorption peak due to O_2 at 60 GHz can be used for secure communication systems.

TABLE 1 IEEE MICROWAVE FREQUENCY BANDS

| Designation | Frequency range (GHz) | Wavelength (cm) |
|---------------|-----------------------|-----------------|
| VHF | 0.1–0.3 | 300.00–100.00 |
| UHF | 0.3–1.0 | 100.00–30.00 |
| L band | 1.0–2.0 | 30.00–15.00 |
| S band | 2.0–4.0 | 15.00–7.50 |
| C band | 4.0–8.0 | 7.50–3.75 |
| X band | 8.0–13.0 | 3.75–2.31 |
| Ku band | 13.0–18.0 | 2.31–1.67 |
| K band | 18.0–28.0 | 1.67–1.07 |
| Ka band | 28.0–40.0 | 1.07–0.75 |
| Millimeter | 30.0–300.0 | 1.00–0.10 |
| Submillimeter | 300.0–3000.0 | 0.10–0.01 |

► 8.2 TUNNEL DIODE

The tunnel diode is associated with the quantum tunneling phenomena.⁵ The tunneling time across the device is very short, permitting its use well into the millimeter-wave region. Because of its mature technology, the tunnel diode is used in special low-power microwave applications, such as local oscillators and frequency-locking circuits.

A tunnel diode consists of a simple p - n junction in which both the p - and n -sides are degenerate (i.e., very heavily doped with impurities). Figure 1 shows a typical static current-voltage characteristic of a tunnel diode under four different bias conditions. The I - V characteristic is the result of two current components: tunneling current and thermal current.

When no voltage is applied to the diode, it is in thermal equilibrium ($V = 0$). Because of the high dopings, the depletion region is very narrow and the tunneling distance d is quite small (5–10 nm). The dopings also cause the Fermi levels to be located within the allowed bands. The amount of degeneracy, qV_p and qV_n , shown at the far left of Fig. 1, is typically 50–200 meV.

When a forward bias is applied, there exists a band of energy states that is occupied on the n side and a corresponding band of energy states that is available and unoccupied on the p side. The electrons can tunnel from the n -side to the p -side. When the applied bias equals approximately $(V_p + V_n)/3$, the tunneling current reaches its peak value I_p and the corresponding voltage is called the peak voltage V_p . When the forward voltage is further increased, there are fewer available unoccupied states on the p -side ($V_p < V < V_v$, where V_v is the valley voltage) and the current decreases. Eventually, the band is “uncrossed,” and at this point the tunneling current can no longer flow. With still further voltage increase, the normal thermal current will flow (for $V > V_v$).

From this discussion we expect that in the forward direction the tunneling current increases from zero to a peak current I_p as the voltage increases. With a further increase in voltage, the current then decreases to zero when $V = V_n + V_p$, where V is the applied forward voltage. The decreasing portion after the peak current in Fig. 1 is the negative differential resistance region. The values of the peak current I_p and the valley current I_v determine the magnitude of the negative resistance. For this reason their ratio I_p/I_v is used as a figure of merit for the tunnel diode.

An empirical form for the I - V characteristics is given by

$$I = I_p \left(\frac{V}{V_p} \right) \exp \left(1 - \frac{V}{V_p} \right) + I_0 \exp \left(\frac{qV}{kT} \right), \quad (1)$$

where the first term is the tunnel current and I_p and V_p are the peak current and peak voltage, respectively, as shown in Fig. 1. The second term is the normal thermal current. The negative differential resistance can be obtained from the first term in Eq. 1:

$$R = \left(\frac{dI}{dV} \right)^{-1} = - \left[\left(\frac{V}{V_p} - 1 \right) \frac{I_p}{V_p} \exp \left(1 - \frac{V}{V_p} \right) \right]^{-1}. \quad (2)$$

Figure 2 shows a comparison of the typical current-voltage characteristics of Ge, GaSb, and GaAs tunnel diodes at room temperature. The current ratios of I_p/I_v are 8:1 for Ge and 12:1 for GaSb and GaAs. Because of its smaller effective mass ($0.042 m_0$) and small bandgap (0.72 eV), the GaSb tunnel diode has the largest negative resistance of the three devices.

► 8.3 IMPATT DIODE

The name IMPATT stands for impact ionization avalanche transit-time. IMPATT diodes employ impact ionization and transit-time properties of semiconductor devices to produce a negative resistance at microwave frequencies. The IMPATT diode is one of the most powerful solid-state sources of microwave power. At present, the IMPATT diode can generate the highest continuous wave (cw) power output of all solid-state devices at millimeter-wave frequencies—above 30 GHz. It is extensively used in radar systems and alarm systems. There is one noteworthy difficulty in IMPATT applications: the noise is high because of random fluctuations of the avalanche multiplication processes.

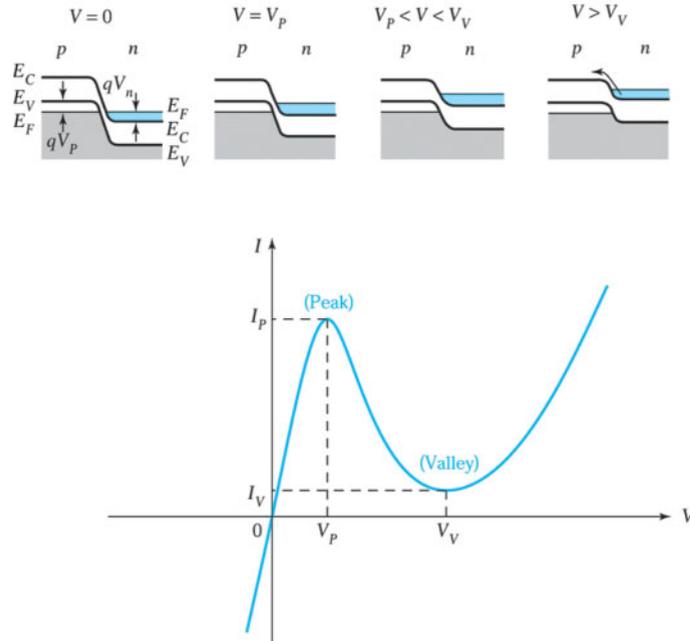


Fig. 1 Static current-voltage characteristics of a typical tunnel diode. I_p and V_p are the peak current and peak voltage, respectively. I_v and V_v are the valley current and valley voltage, respectively. The upper figures show the band diagrams of the device at different bias voltages.

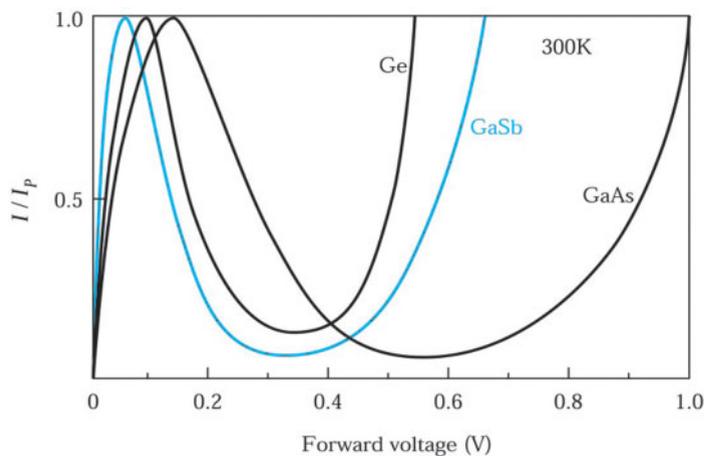


Fig. 2 Typical current-voltage characteristics of Ge, GaSb, and GaAs tunnel diodes at room temperature.

8.3.1 Static Characteristics

The IMPATT diode family includes many different p - n junction and metal-semiconductor devices. The first IMPATT oscillation was obtained from a simple silicon p - n junction diode biased into reverse avalanche breakdown and mounted in a microwave cavity.⁶ Figure 3a shows the doping profile and electric-field distribution at avalanche breakdown of a one-sided abrupt p - n junction. Because of the strong dependence of the ionization rate on the electric field, most of the avalanche multiplication processes occurs in a narrow region near the highest electric field between 0 and x_A (shaded area). x_A is the width of the avalanche region, the distance over which 95% of the contribution to the ionization integrand is obtained.

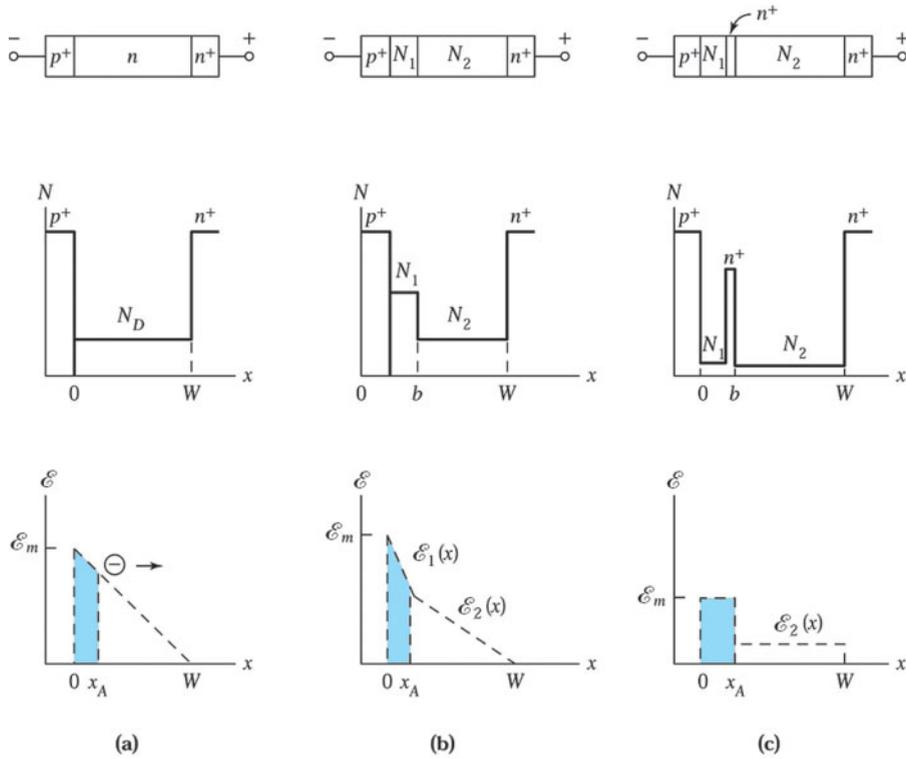


Fig. 3 Doping profiles and electric-field distributions at avalanche breakdown of three single-drift IMPATT diodes: (a) one-sided abrupt p - n junction; (b) hi-lo structure; and (c) lo-hi-lo structure.

Figure 3b shows the hi-lo structure, in which a high doping N_1 region is followed by a lower doping N_2 region. With proper choices of the doping N_1 and its thickness b , the avalanche region can be confined within the N_1 region. Figure 3c is the lo-hi-lo structure in which a “clump” of donor atoms is located at $x = b$. Since a nearly uniform high-field region exists from $x = 0$ to $x = b$, the avalanche region x_A is equal to b , and the maximum field can be much lower than that for a hi-lo structure.

The breakdown voltage V_B (including the built-in potential V_{bi}) is given by the area underneath the electric-field versus distance plot (Fig. 3). For the one-sided abrupt junction (Fig. 3a), V_B is simply $\epsilon_m W/2$. For the hi-lo diode and the lo-hi-lo diode, the breakdown voltages are, respectively:

$$V_B(\text{hi-lo}) = \left(\epsilon_m - \frac{qN_1 b}{2\epsilon_s} \right) b + \frac{1}{2} \left(\epsilon_m - \frac{qN_1 b}{\epsilon_s} \right) (W - b), \quad (3)$$

$$V_B(\text{lo-hi-lo}) = \epsilon_m b + \left(\epsilon_m \frac{qQ}{\epsilon_s} \right) (W - b), \quad (4)$$

where Q in Eq. 4 is the number of impurities/cm² in the clump. The maximum field at breakdown for a hi-lo diode with a given N_1 is the same as the value of the one-sided abrupt junction with the same N_1 . The maximum field of a lo-hi-lo structure can be calculated from the ionization coefficient. These structures are single-drift IMPATT diodes because only one type of charge carriers, electrons, traverses the drift region. If, on the other hand, we form a p^+p - n - n^+ structure, we have a double-drift IMPATT diode in which both electrons and holes participate in device operation over two separate drift regions, i.e., electrons move to the right side and holes move to the left side from the avalanche region. Similar approaches can be used to obtain breakdown voltages for various double-drift diodes.

8.3.2 Dynamic Characteristics

We now use the lo-hi-lo structure, shown in Fig. 3c, to discuss the injection delay and transit-time effect of the IMPATT diode. When a reverse direct current (dc) voltage V_B is applied to the diode so that the critical field for avalanche \mathcal{E}_c is just reached (Fig. 4a), avalanche multiplication will begin. An alternating current (ac) voltage is superimposed onto this dc voltage at $t = 0$. This voltage is shown in Fig 4e. Holes generated in the avalanche region move to the p^+ -region, and electrons enter the drift region. As the applied ac voltage goes positive, more electrons are generated in the avalanche region, as shown by the dotted line in Fig 4b. The electron pulse keeps increasing as long as the electrical field is above \mathcal{E}_c . Therefore, the electron pulse reaches its peak value not at $\pi/2$ when the voltage is maximum, but at π (Fig. 4c). The important consequence is that there is a $\pi/2$ phase delay inherent in the avalanche process itself, that is, the injected-carrier density (electron pulse) lags the ac voltage by 90° .

An additional delay is provided by the drift region. Once the applied voltage drops below V_B ($\pi \leq \omega t \leq 2\pi$), the injected electrons will drift toward the n^+ -contact (Fig. 4d) with a saturation velocity, provided the field across the drift region is sufficiently high.

The situation described above is illustrated by the injected carriers in Fig. 4f. By comparing Figs. 4e and 4f, we note that the peak value of the ac field (or voltage) occurs at $\pi/2$, but the peak of the injected carrier density occurs at π . The injected carriers then traverse the drift region at the saturation velocity, thereby introducing the transit-time delay. The induced external current is also shown in Fig. 4f. Comparing the ac voltage and the external current shows that the diode exhibits a negative resistance characteristic.

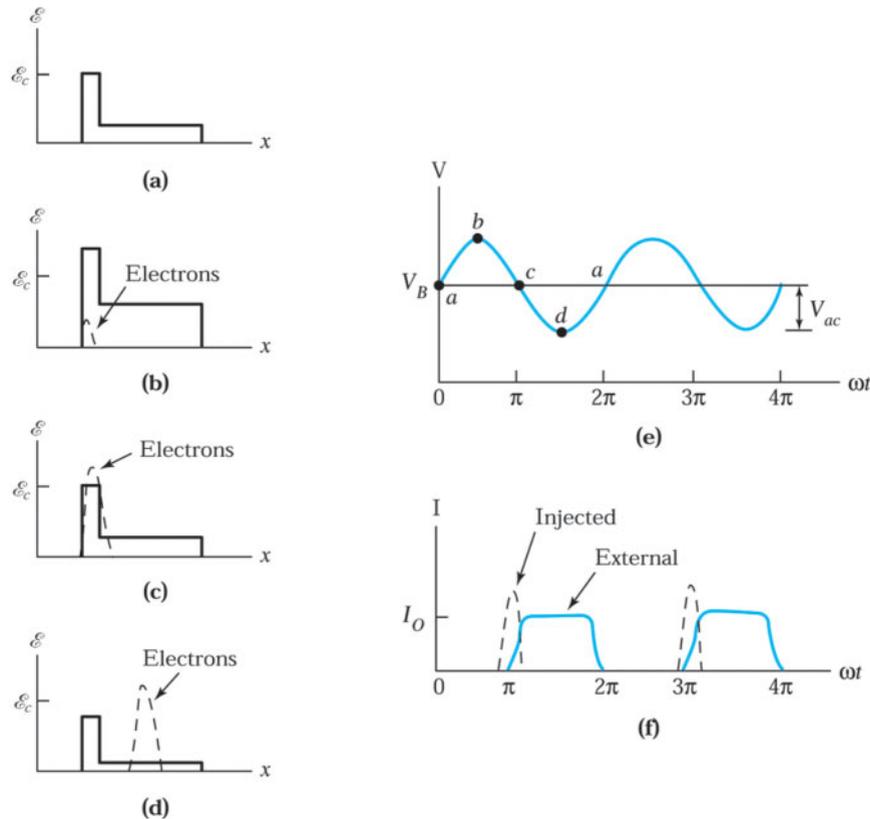


Fig. 4 Field distributions and generated-carrier densities of an IMPATT diode during an ac cycle at four intervals of time (a–d); (e) the ac voltage, and (f) the injected and external current.⁶

The injected carriers (electron pulse) will traverse the length W of the drift region during the negative half-cycle if we choose the transit time to be one-half the oscillation period; that is,

$$\frac{W - x_A}{v_s} = \frac{1}{2} \left(\frac{1}{f} \right) \quad (5)$$

or

$$f = \frac{v_s}{2(W - x_A)}, \quad (6)$$

where v_s is the saturation velocity, which is 10^7 cm/s for silicon at 300 K.

► EXAMPLE 1

Consider a lo-hi-lo silicon IMPATT diode ($p^+ - i - n^+ - i - n^+$) having $b = 1 \mu\text{m}$ and $W = 6 \mu\text{m}$. If the field at breakdown is 3.3×10^5 V/cm, $Q = 2.0 \times 10^{12}$ charges/cm², find the dc breakdown voltage, the field in the drift region, and the operating frequency.

SOLUTION From Eq. 4 we can calculate the breakdown voltage:

$$\begin{aligned} V_B &= 3.3 \times 10^5 \times 10^{-4} + \left(3.3 \times 10^5 - \frac{1.6 \times 10^{-19} \times 2.0 \times 10^{12}}{11.9 \times 8.85 \times 10^{-14}} \right) \times (5 \times 10^{-4}), \\ &= 33 + 13 = 46 \text{ V}. \end{aligned}$$

The field in the drift region is $\frac{13}{5 \times 10^{-4}} = 2.6 \times 10^4$ V/cm.

The drift field is high enough for the injected carriers to maintain their saturation velocity. Therefore

$$f = \frac{v_s}{2(W - x_A)} = \frac{10^7}{2 \times (6 - 1) \times 10^{-4}} = 10^{10} \text{ Hz} = 10 \text{ GHz}. \quad \blacktriangleleft$$

We can also estimate the dc-to-ac power conversion efficiency of the IMPATT diode using Figs. 4e and 4f. The dc power input is the product of the average dc voltage and the average dc current, that is, $V_B(I_0/2)$. The ac power output can be estimated by assuming that the maximum ac voltage swing to be $1/2 V_B$, that is, $V_{ac} = V_B/2$, and the external current is zero between $0 \leq \omega t \leq \pi$ and is I_0 between $\pi \leq \omega t \leq 2\pi$. Therefore, the microwave power-generating efficiency η is

$$\begin{aligned} \eta &= \frac{\text{ac power output}}{\text{dc power input}} = \frac{\int_0^{2\pi} (V_{ac} \sin \omega t) I d(\omega t)}{\left(V_B \frac{I_0}{2} \right) 2\pi} \\ &= \frac{\int_{\pi}^{2\pi} \left(\frac{V_B}{2} \sin \omega t \right) I_0 d(\omega t)}{V_B I_0 \pi} = \frac{1}{\pi} = 32\%. \quad (7) \end{aligned}$$

State-of-the-art IMPATT diodes have cw power capabilities up to 3 W at 30 GHz with over 22% efficiency, up to 1 W at 100 GHz with 10% efficiency, and 50 mW at 250 GHz with 1% efficiency.⁷ The substantial reduction in power and efficiency at higher frequencies is due to difficulties in device fabrication and circuit optimization.

The reduction is also caused by the nonoptimal transit-time delays introduced by the finite time required to transfer energy to the carriers and the tunneling process for the very narrow depletion-layer width.

► 8.4 TRANSFERRED-ELECTRON DEVICES

The transferred-electron effect was first observed in 1963. In the first experiment,⁸ a microwave output was generated when a dc electric field that exceeded a critical threshold value of several thousand volts per centimeter was applied across a short *n*-type sample of GaAs or InP. The transferred-electron device (TED) is an important microwave device. It is used extensively as a local oscillator and power amplifier covering the microwave frequency range from 1 to 150 GHz. The power output and efficiency of TEDs are generally lower than that of IMPATT diodes. However, TEDs have lower noise, lower operating voltages, and relatively easier circuit designs. The TEDs have matured to become important solid-state microwave sources used in detection systems, remote controls, and microwave test instruments.

8.4.1 Negative Differential Resistance

In Chapter 2 we considered the transferred-electron effect, that is, the transfer of conduction electrons from a high-mobility energy valley to low-mobility higher-energy satellite valleys. *N*-type gallium arsenide and *n*-type indium phosphide are the most widely studied and used. Their measured room-temperature velocity field characteristics are shown in Fig. 15 of Chapter 7. Basically, the room-temperature velocity-field characteristics have a region of negative differential resistance (NDR),⁹ as shown in Fig. 5*a*. Also shown are the threshold field \mathcal{E}_T corresponding to the onset of the NDR. The threshold field \mathcal{E}_T is 3.2 kV/cm for gallium arsenide and 10.5 kV/cm for indium phosphide. The peak velocity v_p is about 2.2×10^7 cm/s for gallium arsenide and 2.5×10^7 cm/s for indium phosphide. The maximum negative differential mobility (i.e., $dv/d\mathcal{E}$) is about -2400 cm²/V-s for gallium arsenide and -2000 cm²/V-s for indium phosphide.

For the transferred-electron mechanism to give rise to the NDR, certain requirements must be met. (a) The lattice temperature must be low enough that in the absence of an electric field most of the electrons are in the lower valley (the conduction band minimum); that is, the energy separation between the two valleys $\Delta E > kT$. (b) In the lower valley the electrons must have high mobility and small effective mass, whereas in the upper satellite valleys the electrons must have low mobility and large effective mass. (c) The energy separation between the two valleys must be smaller than the semiconductor bandgap (i.e., $\Delta E < E_g$) so that avalanche breakdown does not begin before the transfer of electrons into the upper valleys.

When a TED is biased in the region of negative resistance with \mathcal{E}_0 as shown in Fig. 5*a*, a momentary space charge and the electrical field distribution become internally unstable. (This is a unique feature of the TED since other negative-resistance devices are stable internally.) The instability in a TED starts with a dipole (also called a domain) which consists of excess electrons (negative charge) and depleted electrons (positive charge),⁹ as shown in Fig. 5*b*. The dipole may arise from any possibility, such as doping inhomogeneity, material defect, or random noise. The dipole is usually formed near the cathode contact, because the largest doping fluctuation and space-charge perturbation exists there. The dipole sets up a higher field for the electrons at that location. This higher field, according to Fig. 5*a*, slows these electrons down relative to the electrons outside the dipole. As a result, the region of excess electrons will grow because the trailing electrons behind the dipole are arriving with a higher velocity. The region of depleted electrons also grows because electrons ahead of the dipole leave with a higher velocity shown in Fig. 5*c*.

Also shown in Fig. 5*c* is that as the dipole grows, the field at that location also increases, but at the expense of the field everywhere else outside the dipole. The field inside the dipole is always above \mathcal{E}_0 , and its carrier velocity decreases monotonically with field. The field outside the dipole is lower than \mathcal{E}_0 , and its carrier velocity goes through the peak value and then decreases monotonically as the field is lowered. When, the field outside the dipole decreases to a certain value \mathcal{E}_s (called sustaining field), the velocities of electrons inside and outside the dipole are the same. At this point the dipole ceases to grow and is said to mature to a domain, usually still near the cathode. The domain then transits from near the cathode to the anode. The terminal-current waveform is shown

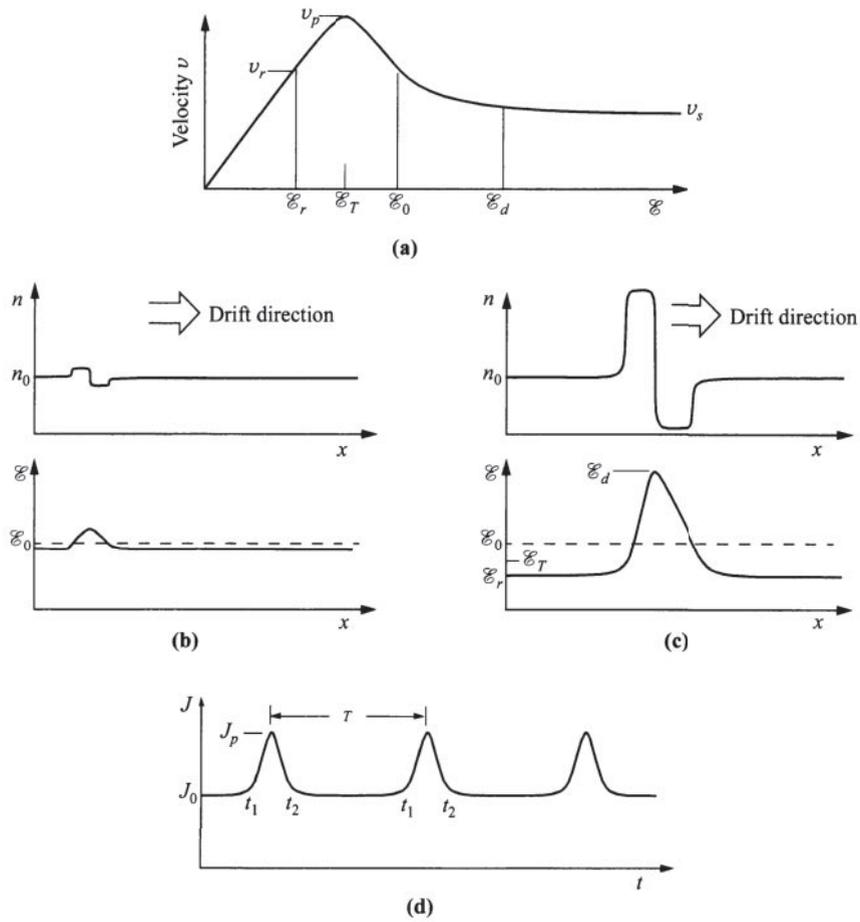


Fig. 5 Demonstration of domain formation. (a) v - \mathcal{E} relationship and some critical point. (b) A small dipole grows to a mature domain. (c) Mature domain. (d) Terminal-current oscillation. Between t_1 and t_2 , the matured domain is annihilated at the anode and another is formed near the cathode.

in Fig. 5d. At t_2 , a domain is formed. At t_1 , the domain reaches the anode. Before another domain is formed, the field throughout the TED jumps to \mathcal{E}_0 . During the formation of a domain (t_2-t_1), the field outside the dipole passes through the value of \mathcal{E}_T where the peak velocity occurs. This causes a peak current. The current pulse width corresponds to the interval between the annihilation of the domain at the anode and the formation of a new domain. The period T corresponding to the transit time of the domain from cathode to anode is L/v , where L is the device length and v is the average velocity. The corresponding frequency is $f = v/L$. The TED can be operated in numerous modes. The above operation of TED is the transit-time domain mode, in which a domain has enough time to mature and transit to the anode.

We now treat the domain formation formally. The one-dimensional continuity equation is given by

$$\frac{\partial n}{\partial t} + \frac{1}{q} \frac{\partial J}{\partial x} = 0 \quad (8)$$

If there is a small local fluctuation of the majority carriers from the uniform equilibrium concentration n_0 , the locally created space charge density is $n - n_0$. Poisson's equation and the current density equation are

$$\frac{\partial \mathcal{E}}{\partial x} = \frac{q(n - n_0)}{\epsilon_s}, \quad (9)$$

$$J = qn_0\bar{\mu}\mathcal{E} + qD\frac{\partial n}{\partial x}, \quad (10)$$

where $\bar{\mu}$ is the average mobility (defined by Eq. 83 in Chapter 2), ϵ_s is the dielectric permittivity, and D is the diffusion constant. Differentiating Eq. 10 with respect to x and inserting Poisson's equation yields

$$\frac{1}{q} \frac{\partial J}{\partial x} = \frac{n - n_0}{\epsilon_s / qn_0\bar{\mu}} + D \frac{\partial^2 n}{\partial x^2}. \quad (11)$$

Substituting this expression into Eq. 8 gives

$$\frac{\partial n}{\partial t} = \frac{n - n_0}{\epsilon_s / qn_0\bar{\mu}} + D \frac{\partial^2 n}{\partial x^2}. \quad (12)$$

We can solve Eq. 12 by separation of variables; that is, let $n(x, t) = n_1(x)n_2(t)$. For the temporal response, the solution of Eq. 12 is

$$n - n_0 = (n - n_0)_0 \exp\left(-\frac{t}{\tau_R}\right), \quad (13)$$

where τ_R is the dielectric relaxation time given by

$$\tau_R = \frac{\epsilon_s}{qn_0\bar{\mu}}. \quad (14)$$

τ_R represents the time constant for the decay of the space charge to neutrality if the mobility $\bar{\mu}$ is positive. However, if the semiconductor exhibits NDR, any charge imbalance will grow with a time constant equal to $|\tau_R|$.

8.4.2 Device Performances

The TEDs require very pure and uniform materials with a minimum of deep impurity levels and traps. Modern TEDs almost always have epitaxial layers on n^+ -substrates deposited by various epitaxial techniques. Typical donor concentrations range from 10^{14} to 10^{16} cm^{-3} and typical device lengths range from a few micrometers to several hundred micrometers. A TED having an epitaxial n -layer on n^+ -substrate and an ohmic n^+ -contact to the cathode electrode is shown in Fig. 6a. Also shown are the energy band diagram at thermal equilibrium and the electric-field distribution when a voltage $V = 3V_T$ is applied to the device where V_T is the product of the threshold field \mathcal{E}_T and the device length L . For such an ohmic contact there is always a low-field region near the cathode. There is no domain formation there due to finite heating time of the lower-valley electrons. The dead zone may be as large as 1 μm , which imposes a constraint on the minimum device length and hence the maximum operating frequency. The field is nonuniform across the device length because the electrical field in the domain grows with distance as seen in Fig. 5b and c.

To improve device performance, we use the two-zone cathode contact instead of the n^+ -ohmic contact. The two-zone cathode contact consists of a high-field zone and an n^+ -zone (Fig. 6b). This configuration is similar to that of a lo-hi-lo IMPATT diode. Electrons are "heated" in the high-field zone and subsequently injected into the active region, which has a uniform field. This structure has been used successfully over a wide temperature range with high efficiency and high power output.

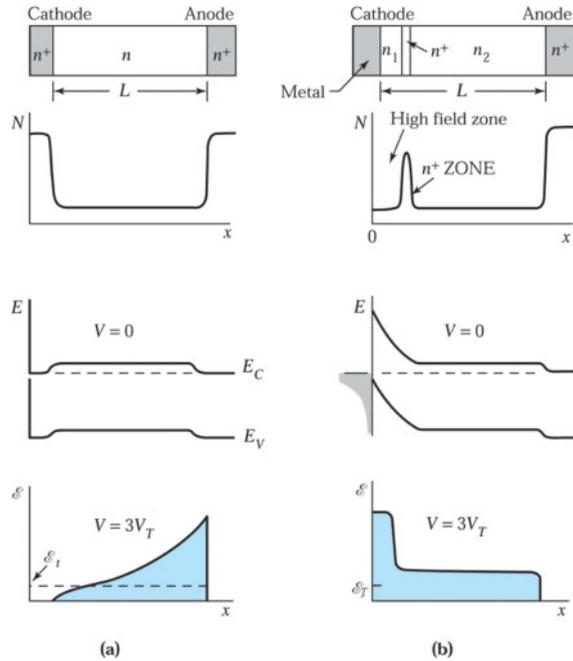


Fig. 6 Two cathode contacts for transferred-electron devices (TEDs). (a) Ohmic contact and (b) two-zone Schottky barrier contact.

We have shown that for a device with NDR, the initial space charge will grow exponentially with time (Eq. 13) and that the time constant is given by Eq. 14:

$$\boxed{|\tau_R| \frac{\epsilon_s}{qn_0|\mu|}}, \quad (15)$$

where μ_- is the negative differential mobility. If Eq. 13 remains valid throughout the entire transit time of the space-charge layer, the maximum growth factor is $\exp(L/v|\tau_R|)$, where L is the length of the active region and v is the average drift velocity of the space-charge layer. For large space-charge growth, this growth factor must be greater than unity, making $L/v|\tau_R| > 1$, or

$$n_0L > \frac{\epsilon_s v}{q|\mu|} \approx 10^{12} \text{ cm}^{-2} \quad (16)$$

for GaAs and InP. A strong space-charge instability is dependent on the condition that enough charge is available in the semiconductor and the device is long enough that the necessary amount of space charge can be built up within the transit time of electrons indicated in Eq. 16. Below this critical n_0L level, the field and carriers are intrinsically stable.

Figure 7 shows a simulated time-dependent behavior of a domain in a gallium arsenide TED 100 μm long with a doping of $5 \times 10^{14} \text{ cm}^{-3}$ ($n_0L = 5 \times 10^{12} \text{ cm}^{-2}$).¹⁰ The time between successive vertical displays of $\mathcal{E}(x,t)$ is $16\tau_R$, where τ_R is the low-field dielectric relaxation time from Eq. 14 ($\tau_R = 1.5 \text{ ps}$ for this device).

State-of-the-art TED diodes have cw power capabilities up to 0.5 W at 30 GHz with 15% efficiency, up to 0.2 W at 100 GHz with 7% efficiency, and 70 mW at 150 GHz with 1% efficiency. The power output of TED is lower than that of IMPATT; however, TED has much lower noise (e.g., 20 dB less at 135 GHz).⁷

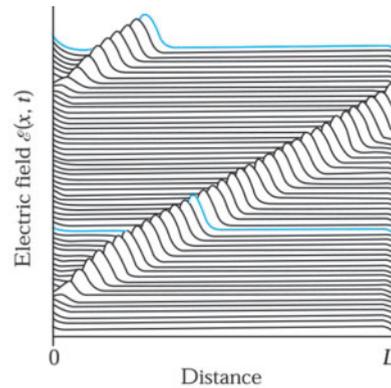


Fig. 7 Numerical simulation of the time-dependent behavior of a cathode-nucleated TED for the transit-time domain mode.¹⁰

► EXAMPLE 2

A GaAs TED is 10 μm long and is operated in the transit-time domain mode. Find the minimum electron density n_0 required and the time between current pulses.

SOLUTION For transit-time domain mode, we require $n_0 L \geq 10^{12} \text{ cm}^{-2}$:

$$n_0 \geq 10^{12} / L = 10^{12} / 10 \times 10^{-4} = 10^{15} \text{ cm}^{-3}.$$

The time between current pulses is the time required for the domain to travel from the cathode to the anode:

$$t = L/v = 10 \times 10^{-4} / 10^7 = 10^{-10} \text{ s} = 0.1 \text{ ns.} \quad \blacktriangleleft$$

The operation modes of a TED depends on five factors: doping concentration and doping uniformity in the device, length of the active region, cathode contact characteristics, type of circuit, and operating bias voltage.¹¹ For example, if no internal domain has built up, the TED is operated in uniform-field mode. It has a uniform electrical field and acts as a regular NDR device. The operating frequency is not restricted to the domain transit time.

If the domain can be quenched before it reaches the anode, the TED is operated in quenched mode. In this mode, domain quenching occurs when the bias voltage is reduced sufficiently below the threshold during an ac cycle. When the bias voltage swings back above the threshold, a new dipole layer is nucleated and the process repeated. The operating frequency can remove the limitation due to the domain transit time. Therefore, the oscillation occurs at the frequency of the resonant circuit.

► 8.5 QUANTUM-EFFECT DEVICES

A quantum-effect device (QED) uses quantum mechanical tunneling to provide controlled carrier transport. In such a device, the active layer thickness is very small, on the order of 10 nm. These small dimensions give rise to a quantum size effect that can alter the band structures and enhance device transport properties. The basic QED is the resonant tunneling diode (RTD), discussed in Section 8.5.1. Many novel current-voltage characteristics can be obtained by combining a RTD with the conventional devices considered in previous chapters. QEDs are of particular importance because they can serve as *functional devices*, that is, they can perform a given circuit function with a greatly reduced number of components.

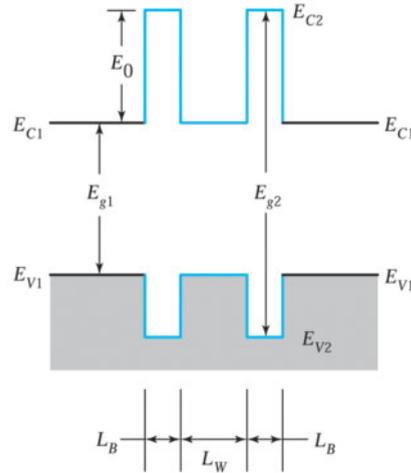


Fig. 8 Band diagram of a resonant-tunneling diode.

8.5.1 Resonant Tunneling Diode

Figure 8 shows the band diagram of a RTD. It has a semiconductor double-barrier structure containing four heterojunctions, a GaAs/AlAs/GaAs/AlAs/GaAs structure, and one quantum well in the conduction band. There are three important device parameters for a RTD—the energy barrier height E_0 , which is the conduction band discontinuity, the energy barrier thickness L_B , and the quantum well thickness L_W .

We now concentrate on the conduction band of a RTD as shown¹² in Fig. 9a. If the well thickness L_W is sufficiently small (on the order of 10 nm or less), a set of discrete energy levels will exist inside the well (such as E_1 , E_2 , E_3 , and E_4 in Fig. 9a). If the barrier thickness L_B is also very small, resonant tunneling will occur. When an incident electron has an energy E that exactly equals one of the discrete energy levels inside the well, it will tunnel through the double barrier with a unity (100%) transmission coefficient.

The transmission coefficient decreases rapidly as the energy E deviates from the discrete energy levels. For example, an electron with an energy 10 meV higher or lower than the level E_1 will result in 10^5 times reduction in the transmission coefficients, as depicted in Fig. 9b. The transmission coefficient can be calculated by solving the one-dimensional Schrödinger equations in the five regions in Fig. 9a (I, II, III, IV, V). Since the wavefunctions and their first derivatives at each potential discontinuity must be continuous, we can obtain the transmission coefficient T_r . Appendix J shows the calculation of the transmission coefficient for the RTD.

The energy levels, E_n , at which the transmission coefficient exhibits its first and second resonant peaks in GaAs/AlAs RTD are shown in Fig. 10a as a function of barrier thickness L_B with the well thickness L_W as a parameter.¹³ It is apparent that E_n is essentially independent of L_B but is dependent on L_W . The calculated width of the peak ΔE_n (i.e., the full width at the half-maximum point of the transmission coefficient where $T_r = 0.5$) is shown in Fig. 10b as function of L_B and L_W . For a given L_W , the width ΔE_n decreases exponentially with L_B .

The cross section of a RTD is shown¹³ in Fig. 11. The alternating GaAs/AlAs layers are grown sequentially by molecular beam epitaxy (MBE) on an n^+ GaAs substrate (the MBE process is discussed in Chapter 11). The barrier thicknesses are 1.7 nm and the well thickness is 4.5 nm. The active regions are defined with ohmic contacts. The top contact is used as a mask to isolate the region under the contact by etching mesas.

The measured current-voltage characteristic of this RTD is shown in Fig. 12. Also shown are the band diagrams for varies dc biases. Note that the I - V curve is similar to that of a tunnel diode (Fig. 1). At thermal equilibrium, $V = 0$, the energy diagram is similar to that in Fig. 9a (here only the lowest energy level E_1 is shown). As we increase the applied voltage, the electrons in the occupied energy states near the Fermi level to the left side of the first barrier tunnel into the quantum well. The electrons subsequently tunnel through the second barrier into the unoccupied states in the right side. Resonance occurs when the energy of the injected electrons becomes approximately equal to the energy level E_1 , where the transmission probability is maximum. This is illustrated by the energy diagram for $V = V_1 = V_p$, where the conduction band edge on the left side is lined up with E_1 .

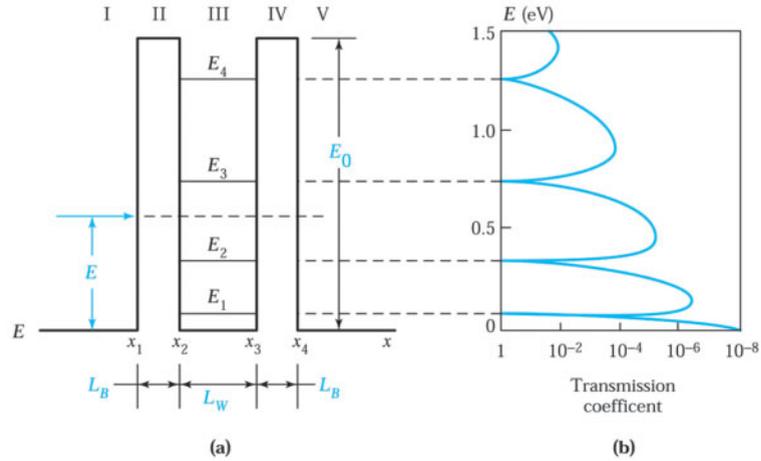


Fig. 9 (a) Schematic illustration of AlAs/GaAs/AlAs double-barrier structure with a 2.5 nm barrier and a 7 nm well. (b) Transmission coefficient versus electron energy for the structure.¹²

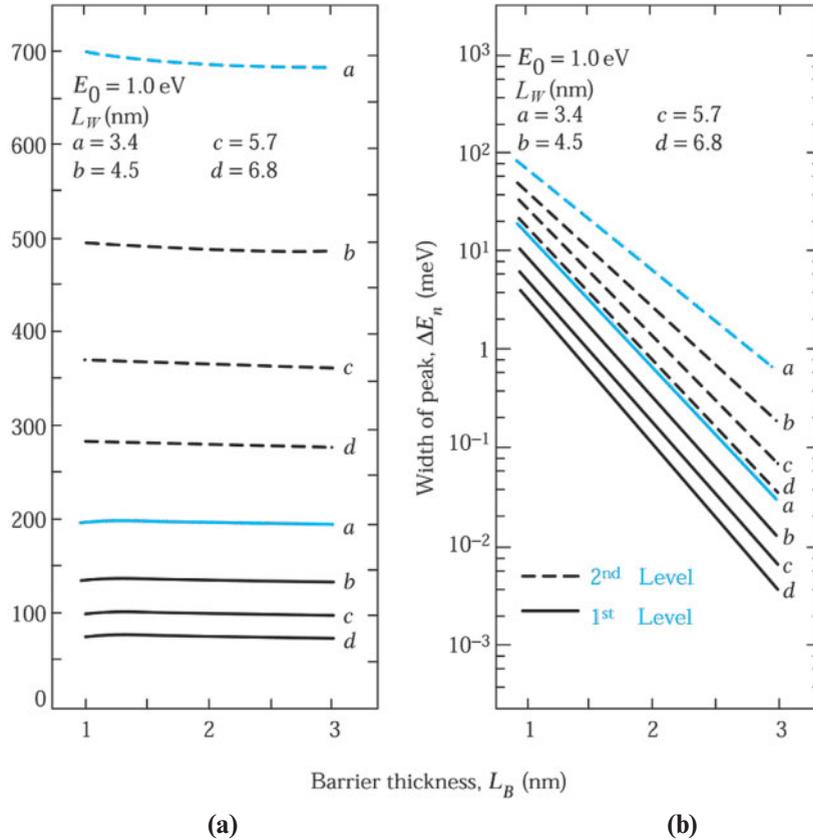


Fig. 10 (a) Calculated energy of electrons at which the transmission coefficient shows the resonant peak in an AlAs/GaAs/AlAs structure as a function of barrier thickness for various well thicknesses. (b) Full width at half maximum of the transmission coefficient versus barrier thickness for the first and second resonant peak.¹³

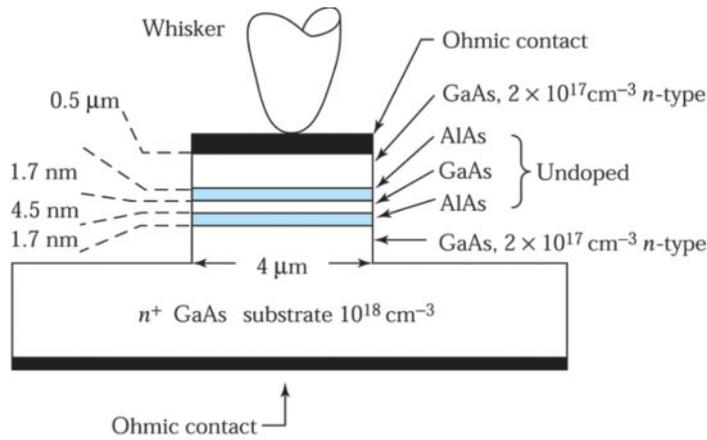


Fig. 11 A mesa-type resonant tunneling diode.¹³

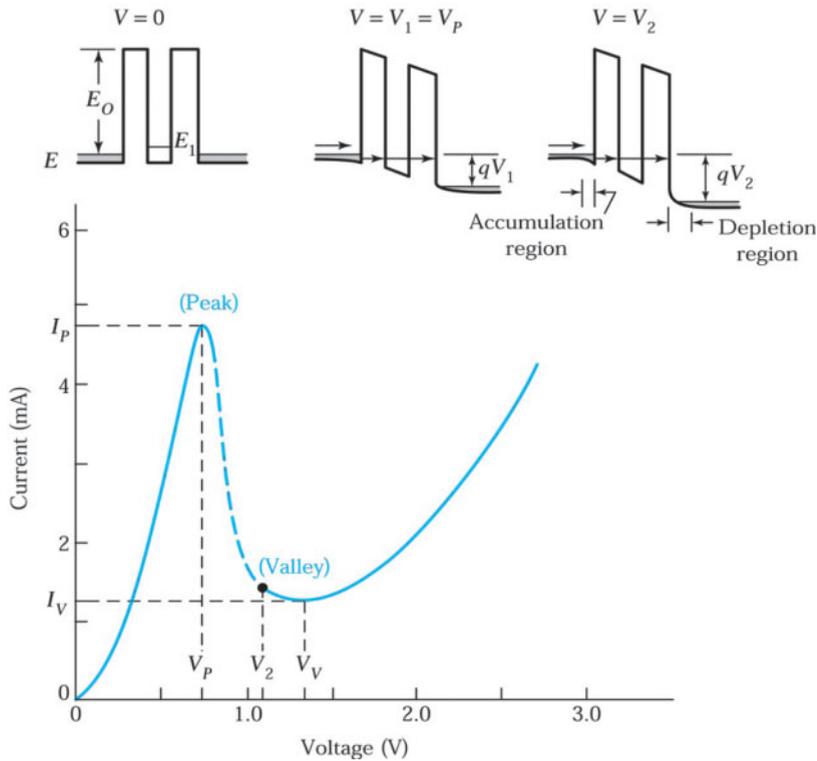


Fig. 12 Measured current–voltage characteristics¹³ of the diode in Fig. 11.

The magnitude of the peak voltage must be at least $2E_1/q$ but is usually larger because of additional voltage drops in the accumulation and depletion regions:

$$V_p > \frac{2E_1}{q}. \tag{17}$$

When the voltage is further increased, that is, at $V = V_2$, the conduction band edge is above E_1 and the number of electrons that can tunnel decreases, resulting in a small current. The valley current I_V is due mainly to the excess current components, such as electrons that tunnel via an upper valley in the barrier. At room temperature and higher, there are other components due to tunneling current associated with either lattice vibrations or impurity atoms. To minimize the valley current, we must improve the quality of the heterojunction interfaces and eliminate impurities in the barrier and well regions. For even higher applied voltages, $V > V_V$, we have the thermionic current component I_{th} , due to electrons injected through higher discrete energy levels in the well or thermionically injected over the barriers. The current I_{th} increases monotonically with increasing voltage similar to that of a tunnel diode. To reduce I_{th} , we should increase the barrier height and design a diode that operates at relatively low bias voltages.

RTDs can be operated at very high frequencies because of their smaller parasitics. In the RTD the main contribution to the capacitance is from the depletion region (refer to the band diagram for $V = V_2$ in Fig. 12). Since the doping density there can be much lower than in a degenerate p - n junction, the depletion capacitance is much smaller. The cutoff frequency for a RTD can reach the THz (10^{12} Hz) range. It can be used in ultra-fast pulse-forming circuits, in THz radiation detection systems, and in oscillators to generate THz signals.

► EXAMPLE 3

Find the ground energy level and the corresponding width of the peak for the RTD in Fig. 11. Compare the peak voltage V_P in Fig. 12 with $2E_1/q$.

SOLUTION From Fig. 10 we find that the ground energy level for $L_w = 4.5$ nm is at 140 meV, and the width of the peak ΔE_1 is about 1 meV. In Fig. 12, V_P is 700 mV, which is larger than 280 mV ($2E_1/q$). The difference (420 mV) is due to voltage drops at the accumulation and the depletion regions. ◀

8.5.2 Unipolar Resonant Tunneling Transistor

A schematic band diagram of a unipolar resonant tunneling transistor¹⁴ is shown in Fig. 13. The structure consists of a resonant tunneling (RT) double barrier placed between the GaAs emitter and base layers. The RT structure was made of a GaAs quantum well (5.6 nm thick) sandwiched between two $\text{Al}_{0.33}\text{Ga}_{0.67}\text{As}$ barriers (5 nm thick). High-energy electrons can be injected from the emitter through the RT into the base region. The electrons are then transported through the n^+ -base region 100 nm thick before being collected at the 300 nm thick $\text{Al}_{0.2}\text{Ga}_{0.8}\text{As}$ collector barrier. The barriers and the quantum well are undoped, whereas the emitter, base, and collector layers are all n -type and doped to $1 \times 10^{18} \text{ cm}^{-3}$.

The operation of the device in the common-emitter configuration with a fixed collector-emitter voltage V_{CE} is shown in the band diagrams of Fig. 13. When the base-emitter voltage V_{BE} is zero (Fig. 13a), there is no electron injection; hence, the emitter and collector currents are zero even with a positive V_{CE} . A peak in the emitter and collector currents occurs when V_{BE} is equal to $2E_1/q$, where E_1 is the energy of the first resonant level in the quantum well (Fig. 13b). With a further increase in V_{BE} , RT is quenched (Fig. 13c) with a corresponding drop in the collector current. The current-voltage characteristic is shown in Fig. 13d. Note that a current peak occurs around $V_{BE} = 0.4$ V. If we connect two inputs A and B to the base terminal (Fig. 13e), the device can perform an exclusive NOR logic function;[§] that is, the output voltage will be high if A and B are both high or both low, otherwise the output voltage will be low. To perform the same function, we need eight conventional MESFETs. Therefore, many quantum-effect devices are useful as functional devices.

[§]Exclusive NOR logic function: when the two inputs are both high or both low, the output is high. Otherwise, the output is low.

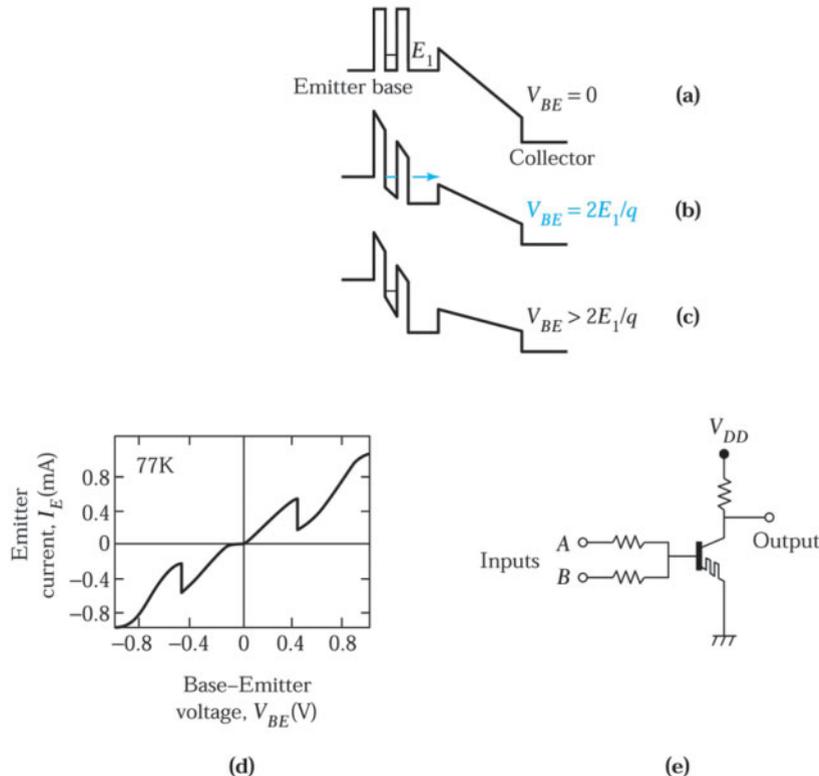


Fig. 13 Band diagrams of a unipolar resonant tunneling (RT) transistor¹³ at (a) $V_{BE} = 0$, (b) $V_{BE} = 2E_1/q$ (maximum RT current), (c) $V_{BE} > 2E_1/q$ (RT quenched), (d) base-emitter current–voltage characteristics measured at 77 K, and (e) an exclusive NOR circuit.

► 8.6 HOT-ELECTRON DEVICES

Hot electrons are electrons with kinetic energies substantially above kT , where k is the Boltzmann constant and T is the lattice temperature. As the dimensions of semiconductor devices shrink and the internal fields rise, a large fraction of carriers in the active regions of the device during its operation are in a state of high kinetic energy. At a given point in time and space the velocity distribution of carriers may be narrowly peaked, in which case one speaks about “ballistic” electron packets. At other times and locations, the electron ensemble can have a broad velocity distribution, similar to a conventional Maxwellian distribution but with an effective electron temperature T_e larger than the lattice temperature T .

Over the years, many hot electron devices have been studied. We now consider two important devices—the hot-electron heterojunction bipolar transistor and the real-space-transfer transistor.

8.6.1 Hot-Electron HBT

Hot-electron injection is made possible in heterojunction bipolar transistors (HBTs) by designing structures with a wider-bandgap emitter,¹⁵ for example, an AlInAs/GaInAs HBT lattice matched to InP. There are several advantages of the hot-electron effect. Electrons are injected by thermionic emission over the emitter-base barrier at an energy $\Delta E_C = 0.5$ eV above the conduction band edge in the p -GaInAs base. Here the purpose of the ballistic injection is to shorten the base traversal time by replacing the relatively slow diffusion motion by faster ballistic propagation.

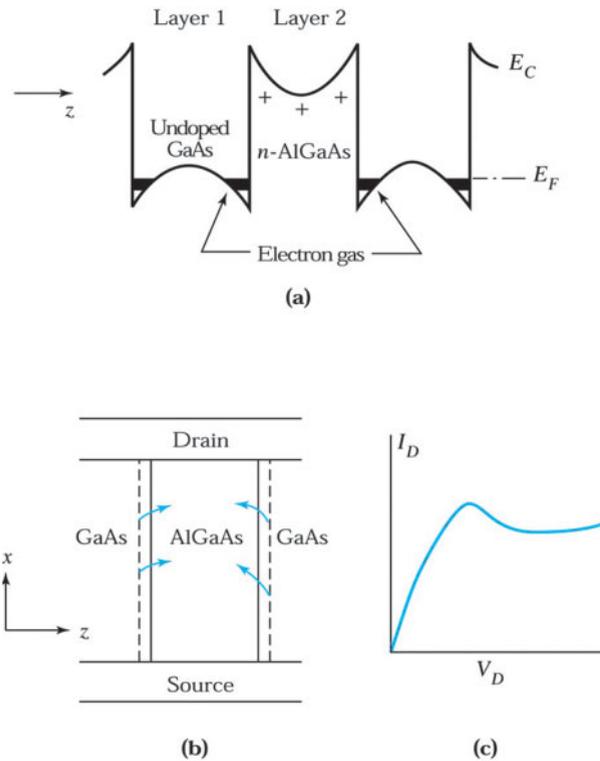


Fig. 14 (a) A heterostructure with alternate GaAs and AlGaAs layers. (b) Electrons, heated by an applied electric field, transfer into the wide-gap layers. (c) If the mobility in layer 2 is lower, the transfer results in a negative differential conductivity.¹⁶

8.6.2 Real-Space-Transfer Transistor

The original real-space transfer (RST) structure, illustrated in Fig. 14a, is a heterostructure with alternate doped wide-gap AlGaAs and undoped narrow-gap GaAs layers. In thermal equilibrium the mobile electrons reside in the undoped GaAs quantum wells and are spatially separated from their parent donors in the AlGaAs layers.¹⁶ If the power input into the structure exceeds the rate of energy loss by the system to the lattice, then the carriers “heat up” and undergo partial transfer into the wide-gap layer, where they may have a different mobility (Fig. 14b). If the mobility in layer 2 is much lower, negative differential resistance will occur in the two-terminal circuit (Fig. 14c). There is a strong analogy to the transferred-electron effect, based on the momentum-space intervalley transfer, whence the name real-space transfer. Two-terminal RST oscillators do not appear to offer any significant advantages over Gunn oscillators because a large mobility ratio between layer 1 and 2 in the RST structure is more difficult to realize than in homogeneous multiple-valley semiconductors. But the possibility of extracting the transferred hot-carrier via a third terminal in an RST transistor makes the RST structure more interesting.

Figure 15 shows a schematic cross-section and the corresponding band diagram of a three-terminal RST transistor (RSTT) implemented in a GaInAs/AlInAs material system.^{17,18} The source and drain contacts are to an undoped high-mobility $\text{Ga}_{0.47}\text{In}_{0.53}\text{As}$ ($E_g = 0.75$ eV with $\mu_n = 13800$ $\text{cm}^2/\text{V}\cdot\text{s}$) channel, and the collector contact is to a doped $\text{Ga}_{0.47}\text{In}_{0.53}\text{As}$ conducting layer. This layer is separated from the channel by a larger-bandgap material. (i.e., an $\text{Al}_{0.48}\text{In}_{0.52}\text{As}$ with $E_g = 1.45$ eV). At $V_D = 0$ an electron density is induced in the source-drain channel by a sufficiently positive collector bias V_C with respect to the grounded source, but no collector current I_C flows because of the AlInAs barrier. As V_D increases, however, a drain current I_D begins to flow and the channel electrons heat

up to some effective temperature T_e (V_D). This electron temperature determines the RST current injected over the AlInAs collector barrier. The injected electrons are swept into the collector by the V_C -induced electric field, giving rise to I_C . Transistor action results from control of the electron temperature T_e in the source-drain channel, which modulates the I_C flowing into the collector electrode.

The drain current I_D and the collector current I_C versus drain voltage V_D at a fixed $V_C = 3.9$ V are shown¹⁹ in Fig. 16. In contrast to the two-terminal device, the RST current is removed from the drain current and leads to very strong NDR in the I_D - V_D curve. On the I_D - V_D characteristics, the RSTT shows pronounced negative differential resistance, with a peak-to-valley ratio that reaches 7000 at 300 K. In the I_C - V_D characteristics, the collector current increases approximately linearly and eventually reaches a saturation value similar to a field-effect transistor.

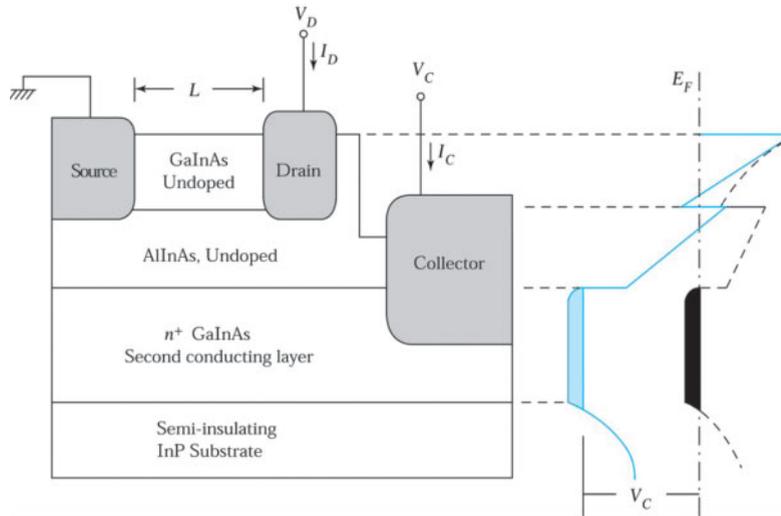


Fig. 15 Cross section and band diagram of a real-space-transfer transistor in a GaInAs/AlInAs material system.^{17,18}

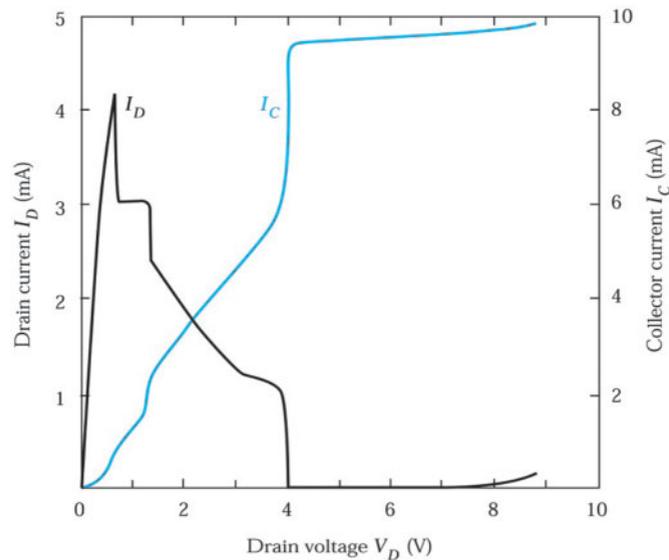


Fig. 16 Experimental real-space-transfer transistor characteristics¹⁹ at $T = 300$ K. Drain current I_D and collector current I_C versus drain voltage V_D at a fixed collector voltage $V_C = 3.9$ V.

The RSTT can be used as a conventional high-speed transistor with high transconductance, $g_m \equiv \partial I_C / \partial V_D$ (at a fixed V_C), and a high cutoff frequency f_T . In addition, the RSTT is another useful functional device for logic circuits. This is because the source and drain contacts of an RSTT are symmetrical. A single device, such as that shown in Fig. 16, can perform an exclusive OR (XOR) logic function* because the collector current I_C flows if the source and drain are at different logic values, regardless of which is “high.”

► SUMMARY

Diodes associated with the tunneling phenomenon (such as the tunnel diode), the avalanche breakdown (IMPATT diode), and momentum-space transfer of electrons (TED) are devices that are used at microwave frequencies. Those two-terminal devices have relatively simple construction and much less parasitic resistance and capacitance than do their three-terminal counterparts. These microwave diodes can operate in the millimeter-wave band (30–300 GHz) and some devices can operate in the submillimeter-wave band (> 300 GHz). Among microwave devices, the IMPATT diode is the most extensively used semiconductor device for millimeter-wave power applications. However, the TED is often used in local oscillators and amplifiers because it has lower noise and can be operated at lower voltages than the IMPATT diode.

We also considered the quantum-effect and hot-electron devices. Quantum effects become important when the device dimensions are reduced to about 10 nm. A key quantum-effect device is the resonant tunneling diode (RTD), which is a heterostructure having a double barrier and a quantum well. If the incoming carrier energy is equal to a discrete energy level in the quantum well, the tunneling probability through the double barrier becomes 100%. This effect is called resonant tunneling. Microwave detectors that operate up to the THz (10^{12} Hz) range have been made using a RTD. By combining RTD with conventional devices we can obtain many novel characteristics. One example is the unipolar resonant tunneling transistor, which can perform a given logic function with a greatly reduced number of components.

HEDs can be classified into two groups—ballistic devices and the RST devices, depending on the type of a hot-electron ensemble employed in the operation. Ballistic devices, such as the hot-electron heterojunction bipolar transistor, have the potential for ultrahigh-speed operation. In ballistic devices, high-kinetic-energy electrons are injected over the emitter-base barrier by thermionic emission. This “ballistic propagation” greatly reduces the transit time through the base. In RST devices, electrons in a narrow-gap semiconductor can gain energy from the input power and undergo transfer into a wide-gap semiconductor, giving rise to a NDR characteristic. These devices such as the RSTT have high transconductance and a high cutoff frequency. RSTTs are also used in logic circuits, where they permit a lower component count for a given function than do other devices.

► REFERENCES

1. S. M. Sze, Ed., *Modern Semiconductor Device Physics*, Wiley, New York, 1998.
2. J. J. Carr, *Microwave and Wireless Communications Technology*, Butterworth-Heinemann, Newton, MA, 1997.
3. L. E. Larson, *RF and Microwave Circuit Design for Wireless Communications*, Artech House, Norwood, MA, 1996.
4. G. R. Thorn, “Advanced Applications and Solid-State Power Sources for Millimeterwave Systems,” *Proc. Soc. Photo-Optic. Inst. Opt. Eng.* (SPIE), **544**, 2 (1985).
5. (a) L. Esaki, “New Phenomenon in Narrow Ge p - n Junction,” *Phys. Rev.*, **109**, 603 (1958); (b) L. Esaki, “Discovery of the Tunnel Diode,” *IEEE Trans. Electron Devices*, **ED-23**, 644 (1976).

*Exclusive OR logic function: when one of the two inputs but not both are high, the output is high.

6. (a) B. C. DeLoach, Jr., "The IMPATT Story," *IEEE Trans. Electron Devices*, **ED-23**, 57 (1976); (b) R. L. Johnston, B. C. DeLoach, Jr., and B. G. Cohen, "A Silicon Diode Oscillator," *Bell Syst. Tech. J.*, **44**, 369 (1965).
7. H. Eisele and G. I. Haddad, "Active Microwave Diodes," in S. M. Sze, Ed., *Modern Semiconductor Device Physics*, Wiley, New York, 1998.
8. J. B. Gunn, "Microwave Oscillation of Current in III-V Semiconductors," *Solid State Comm.*, **1**, 88 (1963).
9. H. Kroemer, "Negative Conductance in Semiconductor," *IEEE Spectr.*, **5**, 47 (1968).
10. M. Shaw, H. L. Grubin, and P. R. Solomon, *The Gunn-Hilsum Effect*, Academic, New York, 1979.
11. S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, 3rd Ed., Wiley Interscience, Hoboken, 2007.
12. M. Tsuchiya, H. Sakaki, and J. Yashino, "Room Temperature Observation of Differential Negative Resistance in AlAs/GaAs/AlAs Resonant Tunneling Diode," *Jpn. J. Appl. Phys.* **24**, L466 (1985).
13. E. R. Brown et al., "High Speed Resonant Tunneling Diodes," *Proc. Soc. Photo-Opt. Inst. Eng.* (SPIE), **943**, 2 (1988).
14. N. Yokoyama et al., "A New Functional Resonant Tunneling Hot Electron Transistor," *Jpn. J. Appl. Phys.*, **24**, L853 (1985).
15. B. Jalali et al., "Near-Ideal Lateral Scaling in Abrupt AlInAs/InGaAs Heterostructure Bipolar Transistor Prepared by Molecular Beam Epitaxy," *Appl. Phys. Lett.*, **54**, 2333 (1989).
16. K. Hess et al., "Negative Differential Resistance Through Real-Space-Electron Transfer," *Appl. Phys. Lett.*, **35**, 469 (1979).
17. S. Luryi, "Hot Electron Transistors," in S. M. Sze, Ed., *High Speed Semiconductor Devices*, Wiley, New York, 1990.
18. S. Luryi and A. Zaslavsky, "Quantum-Effect and Hot-Electron Devices," in S. M. Sze, Ed., *Modern Semiconductor Device Physics*, Wiley, New York, 1998.
19. P. M. Mensz et al., "High Transconductance and Large Peak-to-Valley Ratio of Negative Differential Conductance in Three Terminal InGaAs/InAlAs Real-Space-Transfer Devices," *Appl. Phys. Lett.*, **57**, 2558 (1990).

► **PROBLEMS (* DENOTES DIFFICULT PROBLEMS)**

FOR SECTION 8.2 TUNNEL DIODE

1. Find the depletion-layer capacitance and depletion-layer width at 0.25 V forward bias for a GaAs tunnel diode doped to 10^{19} cm^{-3} on both sides, using the abrupt junction approximation and assuming $V_n = V_p = 0.03 \text{ V}$.
2. The current-voltage characteristic of a GaSb tunnel diode can be expressed by the empirical form of Eq. 1 with $I_p = 10 \text{ mA}$, $V_p = 0.1 \text{ V}$, and $I_0 = 0.1 \text{ nA}$. Find the largest negative differential resistance and the corresponding voltage.

FOR SECTION 8.3 IMPATT DIODE

3. The variation of electric field in the depletion region due to avalanche-generated space charge gives rise to an incremental resistance for abrupt p^+n diode. The incremental resistance is called the space-charge resistance, R_{SC} , and is given by $(1/I) \int_0^W \Delta \mathcal{E} dx$, where $\Delta \mathcal{E}$ is given by

$$\Delta \mathcal{E}(W) = \frac{\int_0^W \rho_s dx}{\epsilon_s} = \frac{IW}{A\epsilon_s v_s}.$$

- (a) Find R_{SC} for a p^+n Si IMPATT diode with $N_D = 10^{15} \text{ cm}^{-3}$, $W = 12 \text{ } \mu\text{m}$, and $A = 5 \times 10^{-4} \text{ cm}^2$. (b) Find the total applied dc voltage for a current density of 10^3 A/cm^2 .
4. A GaAs IMPATT diode is operated at 10 GHz with a dc bias of 100 V and an average biasing current ($I_0/2$) of 100 mA. (a) If the power-generating efficiency is 25% and the thermal resistance of the diode is 10°C/W , find the junction temperature rise above room temperature. (b) If the breakdown voltage increases with temperature at a rate of 60 mV/ $^\circ\text{C}$, find the breakdown voltage of the diode at room temperature.
5. Consider a GaAs single drift lo-hi-lo IMPATT diode shown in Fig. 3c with an avalanche region width (where the electric field is constant) of $0.4 \text{ } \mu\text{m}$ and a total depletion width of $3 \text{ } \mu\text{m}$. The n^+ clump has a charge Q of $1.5 \times 10^{12}/\text{cm}^2$. (a) Find the breakdown voltage of the diode and the maximum field at breakdown. (b) Is the field in the drift region high enough to maintain the velocity saturation of electrons? (c) Find the operating frequency.
- *6. A silicon $n^+p\text{-}\pi\text{-}p^+$ IMPATT diode has a p -layer $3 \text{ } \mu\text{m}$ thick and a π -layer (low-doping p -layer) $9 \text{ } \mu\text{m}$ thick. The biasing voltage must be high enough to cause avalanche break-down in the p -region and velocity saturation in the π region. (a) Find the minimum required biasing voltage and the doping concentration of the p -region. (b) Estimate the transit time of the device.

FOR SECTION 8.4 TRANSFERRED-ELECTRON DEVICES

7. An InP TED is $1 \text{ } \mu\text{m}$ long with a cross-section area of 10^{-4} cm^2 and is operated in transit-time mode. (a) Find the minimum electron density n_0 required for the transit-time mode. (b) Find the time between current pulses. (c) Calculate the power dissipated in the device if it is biased at one-half the threshold.
- *8. (a) Find the effective density of states in the upper valley N_{CU} of the GaAs conduction band. The upper-valley effective mass is $1.2 m_0$. (b) The ratio of electron concentrations between the upper and lower valleys is given by $(N_{CU}/N_{CL}) \exp(-\Delta E/kT_e)$, where N_{CL} is the effective density of states in the lower valley, $\Delta E = 0.31 \text{ eV}$ is the energy difference, and T_e is the effective electron temperature. Find the ratio at $T_e = 300 \text{ K}$. (c) When electrons gain kinetic energies from the electric field, T_e increases. Find the concentration ratio for $T_e = 1500 \text{ K}$.

FOR SECTION 8.5 QUANTUM-EFFECT DEVICES

9. Molecular beam epitaxy interfaces are typically abrupt to within one or two monolayers (one monolayer = 0.28 nm in GaInAs) because of terrace formation in the growth plane. Estimate the energy level broadening for the ground and first excited-electron states of a 15 nm GaInAs quantum well bound by thick AlInAs barriers. (Hint: assume the case of two-monolayer thickness fluctuation and an infinity deep quantum well. The electron effective mass in GaInAs is $0.0427 m_0$.)
10. Find the first excited level and the corresponding width of the peak ΔE_2 for a RTD with AlAs (2 nm)/GaAs (6.78 nm)/AlAs (2 nm). If we want to maintain the same energy level but increase the width ΔE^2 by a factor of 10, what should be the thicknesses of AlAs and GaAs?

Light-Emitting Diodes and Lasers

- ▶ 9.1 RADIATIVE TRANSITIONS AND OPTICAL ABSORPTION
- ▶ 9.2 LIGHT-EMITTING DIODES
- ▶ 9.3 VARIOUS LIGHT-EMITTING DIODES
- ▶ 9.4 SEMICONDUCTOR LASERS
- ▶ SUMMARY

Photonic devices are devices in which the basic particle of light—the photon—plays a major role. We consider four groups of photonic devices: *light-emitting diodes* (LEDs) and *lasers* (light amplification by stimulated emission of radiation), which convert electrical energy to optical energy; *photodetectors*, which electrically detect optical signals; and *solar cells*, which convert optical energy into electrical energy. In this chapter, we focus on LEDs and semiconductor lasers. Photodetectors and solar cells will be discussed in the next chapter.

Specifically, we cover the following topics:

- Basic interactions between a photon and an electron.
- Generation of photons by spontaneous emission for conventional and organic LEDs.
- Generation of photons by stimulated emission for heterostructure lasers.

▶ 9.1 RADIATIVE TRANSITIONS AND OPTICAL ABSORPTION

Figure 1 shows the electromagnetic spectrum of the optical region. The detectable range of light by the human eye extends only from approximately 0.4 μm to 0.7 μm . Also shown are the major color bands from violet to red in the expanded scale. The ultraviolet region includes wavelengths from 0.01 μm to 0.4 μm , and the infrared region extends from 0.7 μm to 1,000 μm . In this chapter, we are primarily interested in the wavelength range from near-ultraviolet ($\sim 0.3 \mu\text{m}$) to near-infrared ($\sim 1.5 \mu\text{m}$).

Figure 1 also shows the photon energy on a separate horizontal scale. To convert the wavelength to photon energy, we use the relationship

$$\lambda = \frac{c}{\nu} = \frac{hc}{h\nu} = \frac{1.24}{h\nu(\text{eV})} \mu\text{m}, \quad (1)$$

where c is the speed of light in vacuum, ν is the frequency of light, h is Planck's constant, and $h\nu$ is the energy of a photon measured in electron volts. For example, a 0.5 μm green light corresponds to a photon energy of 2.48 eV.

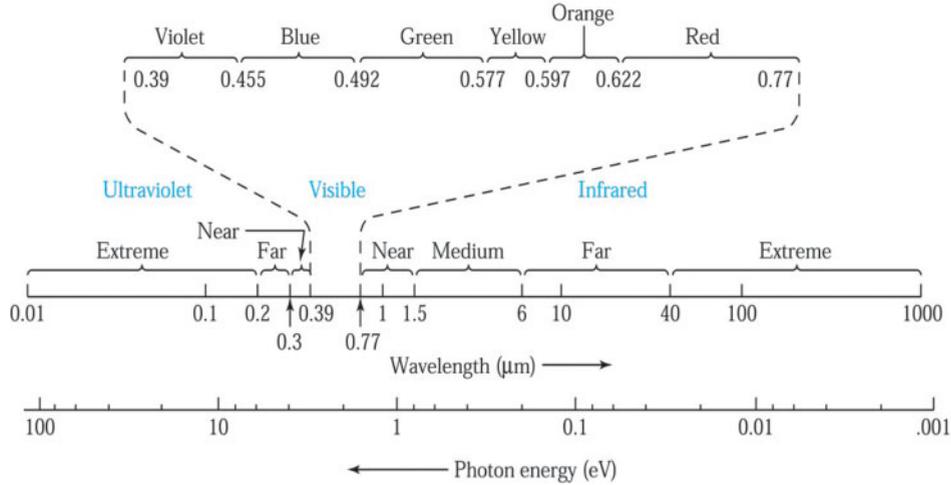


Fig. 1 Chart of the electromagnetic spectrum from the ultraviolet region to the infrared region.

9.1.1 Radiative Transitions

There are basically three processes for interaction between a photon and an electron in a solid: absorption, spontaneous emission, and stimulated emission. We use a simple system to demonstrate these processes.¹ Consider two energy levels E_1 and E_2 of an atom, where E_1 corresponds to the ground state and E_2 corresponds to the excited state (Fig. 2). Any transition between these states involves the emission or absorption of a photon with frequency ν_{12} given by $h\nu_{12} = E_2 - E_1$. At room temperature, most of the atoms in a solid are at the ground state. This situation is disturbed when a photon of energy exactly equal to $h\nu_{12}$ impinges on the system. An atom in state E_1 absorbs the photon and thereby goes to the excited state E_2 . The change in the energy state is the *absorption* process, shown in Fig. 2a. The excited state of the atom is unstable. After a short time, without any external stimulus, it makes a transition to the ground state, giving off a photon of energy $h\nu_{12}$. This process is called *spontaneous emission* (Fig. 2b). When a photon of energy $h\nu_{12}$ impinges on an atom while it is in the excited state (Fig. 2c), the atom can be stimulated to make a transition to the ground state and gives off a photon of energy $h\nu_{12}$, which is in phase with the incident radiation. This process is called *stimulated emission*. The radiation from stimulated emission is monochromatic because each photon has precisely an energy $h\nu_{12}$ and is coherent because all photons emitted are in phase.

The dominant operating process for LEDs is spontaneous emission, for the laser diodes (LDs) it is stimulated emission, and for the photodetectors and the solar cells it is absorption.

Let us assume that the instantaneous populations of E_1 and E_2 are n_1 and n_2 , respectively. Under a thermal equilibrium condition and for $(E_2 - E_1) > 3kT$, the population is given by the Boltzmann distribution:

$$\frac{n_2}{n_1} = e^{-(E_2 - E_1)/kT} = e^{-h\nu_{12}/kT}. \quad (2)$$

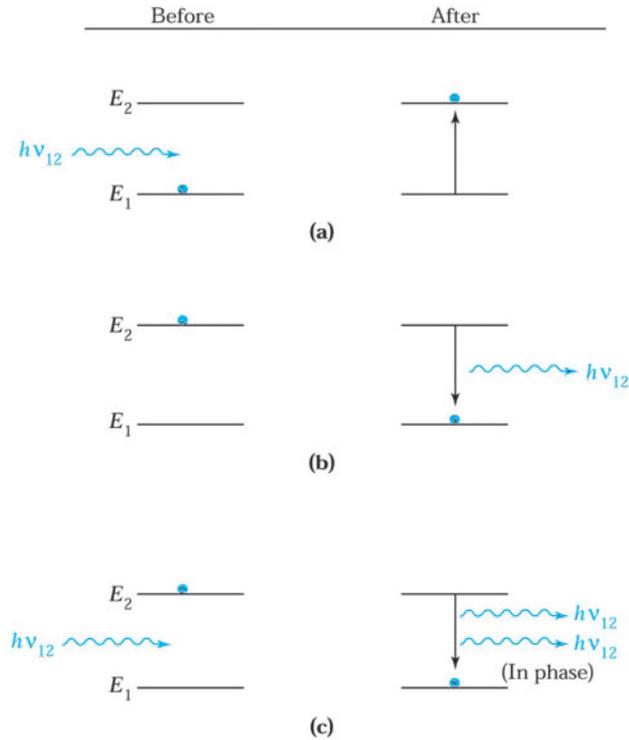


Fig. 2 The three basic transition processes between two energy levels.¹ Black dots indicate the state of the atom. The initial state is at the left; the final state, after the transition, is at the right. (a) Absorption. (b) Spontaneous emission. (c) Stimulated emission.

The negative exponent indicates that n_2 is less than n_1 in thermal equilibrium; that is, most electrons are at the lower energy level.

In steady state, the stimulated-emission rate (i.e., the number of stimulated-emission transitions per unit time) and the spontaneous-emission rate must be balanced by the rate of absorption to maintain the population n_1 and n_2 constant. The stimulated-emission rate is proportional to the photon-field energy density $\rho(h\nu_{12})$, which is the total energy in the radiation field per unit volume per unit frequency. Therefore, the stimulated-emission rate can be written as $B_{21}n_2\rho(h\nu_{12})$, where n_2 is the number of electrons in the upper level and B_{21} is a proportionality constant. The spontaneous-emission rate is proportional only to the population of the upper level and can be written as $A_{21}n_2$, where A_{21} is a constant. The absorption rate is proportional to the electron population at the lower level and to $\rho(h\nu_{12})$; this rate can be written as $B_{12}n_1\rho(h\nu_{12})$, where B_{12} is a proportionality constant. Therefore, we have at steady state

$$\text{stimulated-emission rate} + \text{spontaneous-emission rate} = \text{absorption rate},$$

or

$$B_{21}n_2\rho(h\nu_{12}) + A_{21}n_2 = B_{12}n_1\rho(h\nu_{12}). \quad (3)$$

From Eq. 3 we observe that

$$\frac{\text{stimulated - emission rate}}{\text{spontaneous - emission rate}} = \frac{B_{21}}{A_{21}} \rho(h\nu_{12}). \quad (4)$$

To enhance stimulated emission over spontaneous emission, we must have a very large photon-field energy density $\rho(h\nu_{12})$. To achieve this density, an optical resonant cavity is used to increase the photon field. We also observe from Eq. 3 that

$$\frac{\text{stimulated - emission rate}}{\text{absorption rate}} = \frac{B_{21}}{B_{12}} \left(\frac{n_2}{n_1} \right). \quad (5)$$

If the stimulated emission of photon is to dominate over the absorption of photons, we must have higher electron density in the upper level than in the lower level. This condition is called *population inversion*, since under an equilibrium condition the reverse is true. In Section 9.3 on semiconductor lasers, we consider various ways to have a large photon-field energy density and achieve population inversion, so that the stimulated emission becomes dominant over both spontaneous emission and absorption.

9.1.2 Optical Absorption

Figure 3 shows the basic transitions in a semiconductor. When a semiconductor is illuminated, photons are absorbed to create electron-hole pairs (EHPs), as shown in Fig. 3a, if the photon energy is equal to the bandgap energy, that is, $h\nu$ equals E_g . If $h\nu$ is greater than E_g , an electron-hole pair is generated and, in addition, the excess energy ($h\nu - E_g$) is dissipated as heat, as shown in Fig. 3b. Both processes (Figs. 3a and b) are called *intrinsic transitions* (or band-to-band transitions). On the other hand, for $h\nu$ less than E_g , a photon will be absorbed only if there are available energy states in the forbidden bandgap due to chemical impurities or physical defects, as shown in Fig. 3c. That process is called *extrinsic transition*. This discussion also is generally true for the reverse situation. For example, an electron at the conduction band edge combining with a hole at the valence band edge will result in the emission of a photon with energy equal to that of the bandgap.

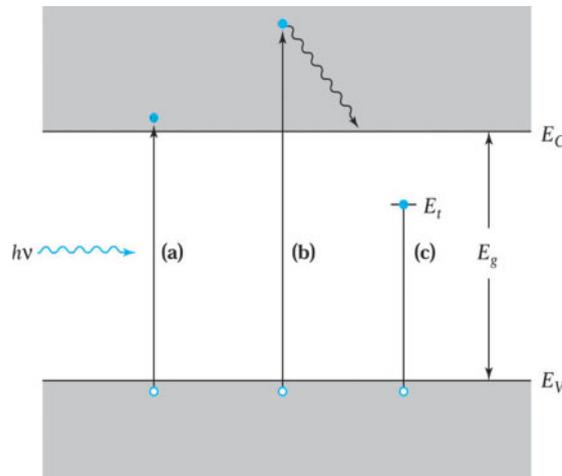


Fig. 3 Optical absorption for (a) $h\nu = E_g$, (b) $h\nu > E_g$, and (c) $h\nu < E_g$.

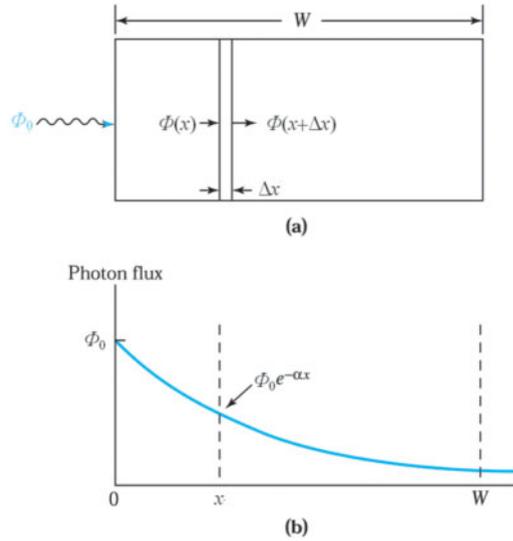


Fig. 4 Optical absorption. (a) Semiconductor under illumination. (b) Exponential decay of photon flux.

Assume that a semiconductor is illuminated from a light source with $h\nu$ greater than E_g and a photon flux of Φ_0 in units of photons per square centimeter per second. As the photon flux travels through the semiconductor, a fraction of the photons absorbed is proportional to the intensity of the flux. Therefore, the number of photons absorbed within an incremental distance Δx (Fig. 4a) is given by $\alpha\Phi(x)\Delta x$, where α is a proportionality constant defined as the absorption coefficient. From the continuity of photon flux as shown in Fig. 4a, we obtain

$$\Phi(x + \Delta x) - \Phi(x) = \frac{d\Phi(x)}{dx} \Delta x = -\alpha\Phi(x)\Delta x$$

or

$$\frac{d\Phi(x)}{dx} = -\alpha\Phi(x). \quad (6)$$

The negative sign indicates a decreasing intensity of the photon flux due to absorption. The solution of Eq. 6 with the boundary condition $\Phi(x) = \Phi_0$, at $x = 0$ is

$$\Phi(x) = \Phi_0 e^{-\alpha x}. \quad (7)$$

The fraction of photon flux that exits from the other end of the semiconductor at $x = W$ (Fig. 4b) is

$$\Phi(W) = \Phi_0 e^{-\alpha W}. \quad (8)$$

The absorption coefficient α is a function of $h\nu$. Figure 5 shows the measured optical absorption coefficient for some important semiconductors that are used for photonic devices.² Also shown is the absorption coefficient for amorphous silicon (dashed curve), which is an important material for solar cells. The absorption coefficient decreases rapidly at the cutoff wavelength λ_c that is

$$\lambda_c = \frac{1.24}{E_g} \mu\text{m} \quad (9)$$

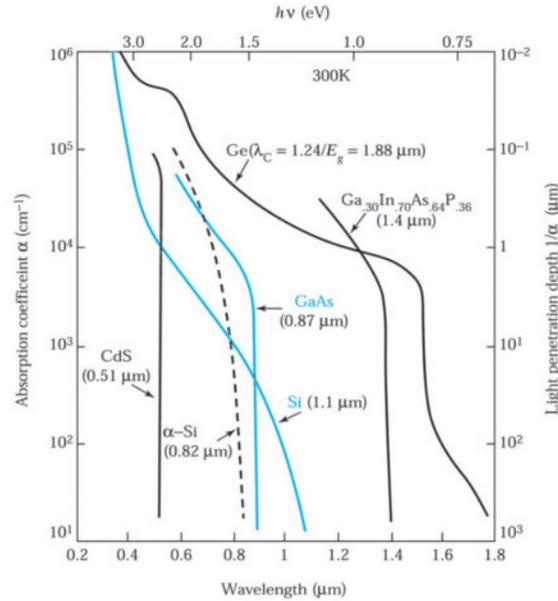


Fig. 5 Optical absorption coefficients for various semiconductor materials.² The value in parentheses is the cutoff wavelength.

because the optical band-to-band absorption becomes negligible for $h\nu < E_g$, or $\lambda > \lambda_c$. From Fig. 5, we notice that 63%, i.e., $(1 - e^{-\alpha W})$ for $\alpha W = 1$, of the optical flux will be absorbed over a distance $W = 1/\alpha$. $1/\alpha$ is called the penetration depth δ , which is also shown in Fig. 5.

The absorption coefficient rises sharply with decreasing wavelength from λ_c for direct bandgap materials GaAs, CdS, $\text{Ga}_{0.30}\text{In}_{0.70}\text{As}_{0.64}\text{P}_{0.36}$ in Fig. 5, because it requires no assistance from lattice vibrations (phonons). For indirect bandgap semiconductors such as Si and Ge, photon absorption requires phonon absorption and emission during the absorption process. The absorption coefficient rises slowly with decreasing wavelength from λ_c . Thus the absorption energy for indirect bandgap semiconductors does not exactly coincide with E_g , but typically it is very close to E_g :

$$h\nu = E_g \pm h\omega \quad (10)$$

where $h\nu$ is the absorption energy, $h\omega$ is the phonon energy.

► EXAMPLE 1

A single-crystal silicon sample $0.25 \mu\text{m}$ thick is illuminated with a monochromatic (single-frequency) light having an $h\nu$ of 3 eV. The incident power is 10 mW. Find the total energy absorbed by the semiconductor per second, the rate of excess thermal energy dissipated to the lattice, and the number of photons per second given off from recombination by intrinsic transitions.

SOLUTION From Fig. 5, the absorption coefficient α is $4 \times 10^4 \text{ cm}^{-1}$. The energy absorbed per second is

$$\begin{aligned} h\nu\Phi_0(1 - e^{-\alpha W}) &= 10^{-2} \left[1 - \exp\left(-4 \times 10^4 \times 0.25 \times 10^{-4}\right) \right] \\ &= 0.0063 \text{ J/s} = 6.3 \text{ mW}. \end{aligned}$$

The portion of each photon's energy that is converted to heat is

$$\frac{h\nu - E_g}{h\nu} = \frac{3 - 1.12}{3} = 62\%.$$

Therefore, the amount of energy dissipated per second to the lattice is

$$62\% \times 6.3 = 3.9 \text{ mW}.$$

Since the recombination radiation accounts for 2.4 mW (i.e., 6.3 mW – 3.9 mW) at 1.12 eV/photon, the number of photons per second from recombination is

$$\frac{2.4 \times 10^{-3}}{1.6 \times 10^{-19} \times 1.12} = 1.3 \times 10^{16} \text{ photons / s.}$$

▶ 9.2 LIGHT-EMITTING DIODES

Light-emitting diodes (LEDs) are p - n junctions that can emit spontaneous radiation in ultraviolet, visible, or infrared regions. The visible LED has a multitude of applications as an information link between electronic instruments and their users. The infrared LED is useful in opto-isolators and for optical-fiber communication.

9.2.1 Structure of LED

The basic structure of an LED is a p - n junction. Under forward bias, electrons are injected from the n -side and holes from the p -side as shown in Fig. 6a. The built-in potential of the junction is lowered by an amount equal to the applied potential V , and the injected carriers can pass across the junction where they become excess minority carriers. In the vicinity of the junction, the excess of carriers is more than the equilibrium value ($pn > n_i^2$), and the recombination will take place, as shown in Fig. 6b. However, if a double heterojunction design is utilized, the LED efficiency can be much improved. Figure 6c shows the central material, which is bound by layers with a higher energy gap. Excess carriers of both types are injected and confined at the same space to produce light. The number of excess carriers in the central region can be significantly increased. The radiative recombination lifetime is shortened due to higher EHP concentrations, and more efficient radiative recombination is obtained. In this configuration, the central layer is usually undoped and bound by layers of opposite types. This double-heterojunction design yields a much higher efficiency and is the preferred approach.

Furthermore, if the thickness of the central active layer is reduced to the range of 10 nm or less, a quantum well is formed. A quantum well is a potential well that confines carriers, which were originally free to move in three dimensions, to two dimensions. Two-dimensional carrier densities become sharper at the band edge, as discussed later and in Appendix H. The carrier densities can be pushed to higher levels and can result in higher recombination efficiency. Another advantage of a thin active layer is that it can accommodate higher level of lattice mismatch in epitaxial growth. Epitaxial growth will be introduced in Chapter 11.

9.2.2 Optical characteristics of the LED

Band-to-band Recombination

The recombination of electron and hole produces a photon with energy $h\nu$ nearly equal to the band gap. Because the largest electron concentration in the conduction band is at an energy $kT/2$ above E_C , similarly the most probable energy for a hole in the valence band has an energy $kT/2$ below E_V . The photon energy is approximately

$$h\nu = E_g + kT. \quad (11)$$

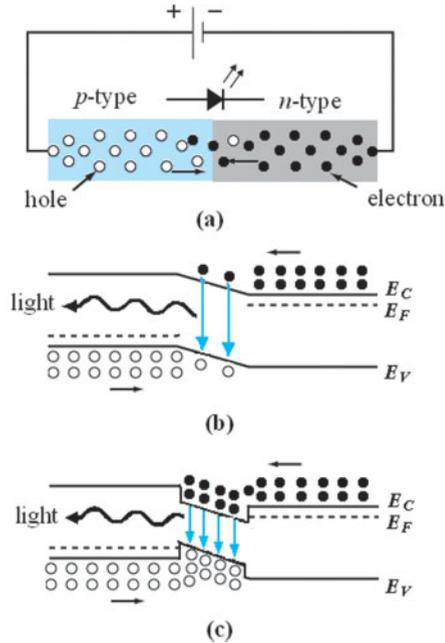


Fig. 6 (a) Under forward bias of a p - n junction, electrons injected from n -side recombine with holes injected from p -side. (b) Recombination taking place in the vicinity of the junction. (c) Higher carrier densities confined in a double heterojunction.

For practical purposes we can ignore the kT term and

$$h\nu = E_g. \quad (11a)$$

Spectral Width

The spectral width is given by the full width at half maximum (FWHM) intensity. The light spectrum of an LED shows a peak at the wavelength λ_m given by Eq. (11). A simple differentiation of Eq. (1) with respect to λ shows that there is a spread of wavelengths $\Delta\lambda$ associated with a spread of energy ΔE :

$$\Delta\lambda \approx \frac{1}{hc} \lambda^2 \Delta E. \quad (12)$$

ΔE is given by kT from Eq. (11). The spectral width on either side of the wavelength peak thus has a dependence on λ^2 and T .³ The spectral width is given by $2\Delta\lambda$. The FWHM becomes larger as the wavelength is increased from visible to infrared. For example, at $\lambda_m = 0.55 \mu\text{m}$ (green), FWHM is about 20 nm, but at $1.3 \mu\text{m}$ (infrared), FWHM is over 120 nm.

Frequency Response

The electrical input signal is generally modulated at high frequencies. This signal causes direct modulation of the injected current in an LED. Parasitic elements such as the depletion-layer capacitance and series resistance can cause a delay of carrier injection into the junction and a delay in the light output. The ultimate limit on how fast one can vary the light output depends on the carrier lifetime, which is determined by various recombination processes, such as the surface recombination discussed in Chapter 2. If the current is modulated at an angular frequency ω , the light output $P(\omega)$ is given by

$$P(\omega) = \frac{P(0)}{\sqrt{1 + (\omega\tau)^2}} \quad (13)$$

where $P(0)$ is the light output at $\omega = 0$ and τ is the overall carrier lifetime. The modulation band-width Δf is defined as the frequency at which the light output is reduced to $1/\sqrt{2}$ from that at $\omega = 0$, that is,

$$\Delta f \equiv \frac{\Delta\omega}{2\pi} = \frac{1}{2\pi\tau}. \quad (14)$$

The overall carrier lifetime τ is related to the radiative (τ_r) and nonradiative (τ_{nr}) lifetimes:

$$\frac{1}{\tau} = \frac{1}{\tau_r} + \frac{1}{\tau_{nr}}. \quad (15)$$

τ approaches τ_r when $\tau_r \ll \tau_{nr}$. τ_r decreases as the doping in the active layer is increased and Δf becomes larger. To increase speed, an increase of doping concentration in the middle active layer of the heterostructure is desirable. The frequency response determines the maximum frequency at which the LED can be turned on and off, and thus the maximum transmission rate of data.

► EXAMPLE 2

Calculate the modulation bandwidth of a GaAs-based LED with $\tau = 500$ ps.

SOLUTION From Eq. 14,

$$\Delta f = \frac{1}{2\pi \cdot 500 \cdot 10^{-12}} = 318 \text{ MHz.} \quad \blacktriangleleft$$

9.2.3 Quantum Efficiency

Internal Quantum Efficiency

For a given input power, the radiative recombination processes are in competition with the nonradiative ones. The band-to-band transition and transitions via traps can be either radiative or nonradiative. Examples of nonradiative band-to-band recombination are those in indirect-bandgap semiconductors. Examples of radiative recombination via traps are those via isoelectronic levels, which will be discussed later.

The internal quantum efficiency η_{in} is the efficiency of converting injected carriers to emitted photons, defined as

$$\eta_{in} = \frac{\text{number of photons emitted internally}}{\text{number of carriers passing the junction}}. \quad (16)$$

This can be related to the fraction of the injected carriers that recombine radiatively to the total recombination rate, and may also be written in terms of their lifetimes as

$$\eta_{in} = \frac{R_r}{R_r + R_{nr}} = \frac{\tau_{nr}}{\tau_{nr} + \tau_r}. \quad (17)$$

Here R_r and R_{nr} are the radiative and nonradiative recombination rates, respectively. It is evident that the radiative lifetime τ_r needs to be small to yield high internal quantum efficiency. For low-level injection, the radiative recombination rate in the p -side of the junction is given by

$$R_r = R_{ec}np \sim R_{ec}\Delta nN_A, \quad (18)$$

where R_{ec} is the recombination coefficient and Δn is the excess carrier density, which is much larger than the minority carrier density in equilibrium, $\Delta n \gg n_{p0}$. R_{ec} is a function of the band structure and temperature. Its value is $\sim 10^{-10}$ cm³/s for direct bandgap materials, and much smaller for indirect bandgap materials ($R_{ec} \sim 10^{-15}$ cm³/s).

For low-level injection ($\Delta n < p_{p0}$), the radiative lifetime τ_r is related to the recombination coefficient by

$$\tau_r = \frac{\Delta n}{R_r} = \frac{1}{R_{ec}N_A}. \quad (19)$$

For high-level injection, however, τ_r would decrease with increasing Δn due to higher probability of carriers recombination. So in double-heterostructure LEDs, carrier confinement increases Δn and τ_r is reduced to improve the internal quantum efficiency as in Eq. (17).

The nonradiative lifetime is usually ascribed to traps or recombination centers of density N_t ,

$$\tau_{nr} = \frac{1}{\sigma v_{th} N_t}, \quad (20)$$

where σ is the capture cross section and v_{th} is the average thermal velocity.

External Quantum Efficiency

Obviously, for LED applications, what matters is the light emitted external to the device. The parameter to measure the efficiency of the light emitted externally is the optical efficiency η_{op} , sometimes called the extraction efficiency. The external quantum efficiency is defined as

$$\eta_{ex} = \frac{\text{number of photons emitted externally}}{\text{number of carriers passing the junction}} = \eta_{in}\eta_{op}. \quad (21)$$

Basically, there are several main loss mechanisms that reduce the optical efficiency. We focus on the device optical paths and optical interfaces.

1. Absorption within the LED material: The magnitude of the loss is related to the absorption coefficient for a given photon wavelength, as discussed in Section 9.1. Absorption can be minimized by placing the junction closer to the emitting surface.
2. Absorption in the substrate: The direct-bandgap GaAsP LED, which emits red light, fabricated on a GaAs substrate, is shown in Fig. 7a. The indirect bandgap GaAsP LED with higher bandgap energy, which emits orange, yellow, or green light fabricated on a GaP substrate, is shown in Fig. 7b. A graded-alloy GaAs_{1-y}P_y layer is grown epitaxially to minimize the nonradiative centers that result from lattice mismatch at the interface. The absorption loss for a GaAsP red LED on a GaAs substrate is large since the substrate is opaque to light and it absorbs about 85% of the photons emitted at the junction, as shown in Fig. 7a. For orange, yellow, or green GaAsP LEDs on a GaP transparent substrate, photons emitted downward can be reflected back with only about 25% absorption by the bottom metal contact. The efficiency can be significantly improved, as shown in Fig. 7b. Of course, the absorption in the substrate can be less with a thinner substrate. After LED fabrication, the substrate is usually ground to about 100 μm to enhance extraction efficiency. However, too thin a substrate will lose mechanical strength.

3. Fresnel reflection loss: For normal light incidence from semiconductor to air, the direction of the optical path is not changed. But it suffers from the Fresnel loss with a reflection coefficient associated with the different refractive indices shown as the optical path A in Fig. 7a:

$$R = (\bar{n}_1 - \bar{n}_2)^2 / (\bar{n}_1 + \bar{n}_2)^2, \tag{22}$$

where \bar{n}_1 and \bar{n}_2 are the refractive indices of semiconductor and outside medium (usually air $\bar{n}_1 = 1$). This optical loss can be minimized by an anti-reflection coating on the LED surface.

4. Total internal reflection loss: The incident light at an angle greater than the critical angle θ_c defined by Snell's law will be totally reflected back to the semiconductor shown as the optical path B in Fig. 7a:

$$\sin \theta_c = \frac{\bar{n}_1}{\bar{n}_2}, \tag{23}$$

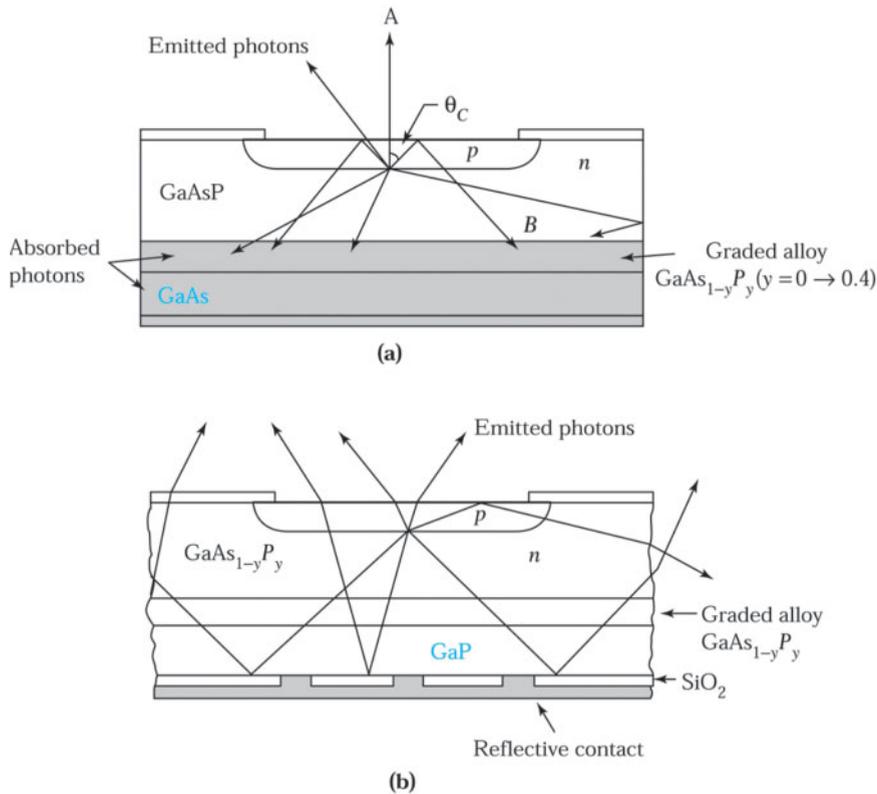


Fig. 7 Basic structure of a flat-diode LED and the effects of (a) an opaque substrate (GaAs_{1-y}P_y) and (b) a transparent substrate (GaP) on photons emitted at the p-n junction.⁴

where the light passes from a medium with a refractive index of \bar{n}_2 , such as GaAs with $\bar{n}_2 = 3.66$ at $\lambda = 0.8 \mu\text{m}$, to a medium of \bar{n}_1 , such as air with $\bar{n}_1 = 1$. For GaAs, the critical angle is about 16° and for GaP with $\bar{n}_2 = 3.45$ at $\lambda = 0.8 \mu\text{m}$, the critical angle is about 17° . The total internal reflection can be minimized by surface roughness.⁵

The forward current-voltage characteristics of a LED are similar to those of the GaAs p - n junction discussed in Chapter 3. At low forward voltages, the diode current is dominated by the nonradiative recombination current due mainly to surface recombination near the perimeter of LED chip. At higher forward voltages, the diode current is dominated by the radiative diffusion current. At even higher voltages, the series resistance will limit the diode current. The total diode current can be written as

$$I = I_d \exp\left[\frac{q(V - IR_s)}{KT}\right] + I_r \exp\left[\frac{q(V - IR_s)}{2KT}\right], \quad (24)$$

where R_s is the device series resistance and I_d and I_r are the saturation currents due to diffusion and recombination, respectively. To increase the output power of LED, we must reduce I_r and R_s .

► 9.3 VARIOUS LIGHT-EMITTING DIODES

Light-emitting diodes (LEDs) are p - n junctions that can emit spontaneous radiation in ultraviolet, visible, or infrared regions. The visible LED has a multitude of applications as an information link between electronic instruments and their users. The white LED has become a key component in backlight source for the liquid-crystal flat-panel display and street lamps. It has a potential to replace the conventional light sources for solid-state lighting applications when the costs of blue, green and red LEDs, especially blue LEDs, become competitive. The infrared LED is useful in opto-isolators, optical-fiber communication and health care applications.

9.3.1 Visible LEDs

Figure 8 shows the relative eye response as a function of wavelength (or the corresponding photon energy). The maximum sensitivity of the eye is at 555 nm. The eye response falls to nearly zero at the extremes of the visible spectrum at about 400 and 700 nm. For normal vision at the peak response of the eye, 1 W of radiant energy is equivalent to 683 lumen.

Since the eye is sensitive only to light with a photon energy $h\nu$ equal to or greater than 1.8 eV ($< 700 \text{ nm}$), semiconductors of interest must have an energy bandgap larger than this limit. Figure 8 also shows the bandgaps of various semiconductors. Table 1 lists the semiconductors used to produce light in the visible and infrared parts of the spectrum. Among all the semiconductors shown, the most important materials for visible LEDs are the alloy $\text{GaAs}_{1-y}\text{P}_y$ and $\text{Ga}_x\text{In}_{1-x}\text{N}$ III-V compound systems. An alloy III-V compound is formed when more than one group III element is distributed randomly on group III lattice sites (e.g., gallium sites) or more than one group V element is distributed randomly on group V lattice sites (e.g., arsenic sites). The notation used is $\text{A}_x\text{B}_{1-x}\text{C}$ or $\text{AC}_{1-y}\text{D}_y$ for ternary (three elements) compounds and $\text{A}_x\text{B}_{1-x}\text{C}_y\text{D}_{1-y}$ for quaternary (four elements) compounds, where A and B are the group III elements, C and D are the group V elements, and x and y are the mole fractions, that is, the ratios of the number of atoms of a given species to the total number of group III or group V atoms in the alloy compound.

Figure 9a shows the energy gap for $\text{GaAs}_{1-y}\text{P}_y$ as a function of the mole fraction y . For $0 < y < 0.45$, the bandgap is direct and increases from $E_g = 1.424 \text{ eV}$ at $y = 0$ to $E_g = 1.977 \text{ eV}$ at $y = 0.45$. For $y > 0.45$, the bandgap is indirect. Figure 9b shows the corresponding energy-momentum plots for selected alloy compositions.⁶ As indicated, the conduction band has two minima. The one along $p = 0$ is the direct minimum and the one along $p = p_{\text{max}}$ is the indirect minimum. Electrons in the direct minimum of the conduction band and holes at the top of the valence band have equal momenta ($p = 0$). But electrons in the indirect minimum of the conduction band and holes at the top of the valence band have different momenta. The radiative transition mechanisms are found predominantly in direct-bandgap semiconductors, such as GaAs and $\text{GaAs}_{1-y}\text{P}_y$ ($y < 0.45$).

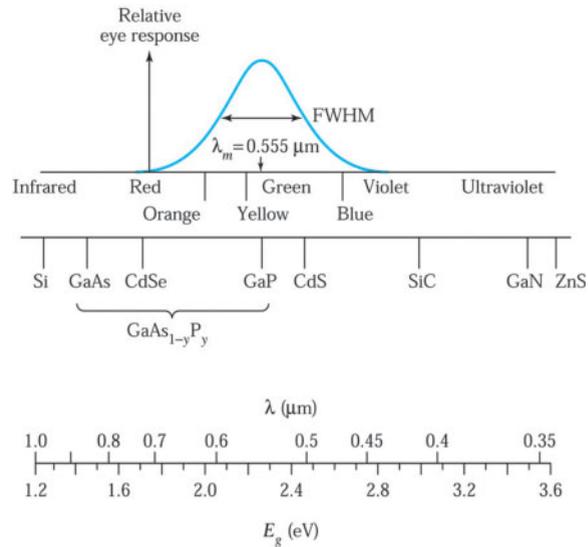


Fig. 8 Semiconductors of interest for visible LEDs, including the relative luminosity function of the human eye.

TABLE 1 COMMON III-V MATERIALS USED TO PRODUCE LEDs AND THEIR EMISSION WAVELENGTHS

| Material | Wavelength (nm) |
|---------------------------------------|-----------------|
| InAsSbP/InAs | 4200 |
| InAs | 3800 |
| GaInAsP/GaSb | 2000 |
| GaSb | 1800 |
| $Ga_xIn_{1-x}As_{1-y}P_y$ | 1100-1600 |
| $Ga_{0.47}In_{0.53}As$ | 1550 |
| $Ga_{0.27}In_{0.73}As_{0.63}P_{0.37}$ | 1300 |
| GaAs:Er,InP:Er | 1540 |
| Si:C | 1300 |
| GaAs:Yb,InP:Yb | 1000 |
| $Al_xGa_{1-x}As:Si$ | 650-940 |
| GaAs:Si | 940 |
| $Al_{0.11}Ga_{0.89}As:Si$ | 830 |
| $Al_{0.4}Ga_{0.6}As:Si$ | 650 |
| $GaAs_{0.6}P_{0.4}$ | 660 |
| $GaAs_{0.4}P_{0.6}$ | 620 |
| $GaAs_{0.15}P_{0.85}$ | 590 |
| $(Al_xGa_{1-x})_{0.5}In_{0.5}P$ | 655 |
| GaP | 690 |
| GaP:N | 550-570 |
| $Ga_xIn_{1-x}N$ | 340,430,590 |
| SiC | 400-460 |
| BN | 260,310,490 |

However, for $\text{GaAs}_{1-y}\text{P}_y$ with y greater than 0.45 and GaP , which are indirect-bandgap semiconductors, the probability of radiative transitions is very small, since lattice interactions or other scattering agents must participate in the process to conserve momentum. Therefore, for indirect-bandgap semiconductors, special recombination centers are incorporated to enhance the radiative processes. Incorporating nitrogen into the crystal lattice can form efficient radiative recombination centers in $\text{GaAs}_{1-y}\text{P}_y$.

When nitrogen is introduced, it replaces phosphorus atoms in the lattice sites. Nitrogen and phosphorus belong to the same group in the period table. The outer electronic structure of nitrogen is similar to that of phosphorus, but the electronic core structures of these atoms are different. This difference introduces a trap level below the conduction band. The trapped electron subsequently can attract a hole and recombine to give a radiative emission with the energy of 50 meV smaller than the band gap. Nitrogen serves as a recombination center but cannot contribute extra carriers, and therefore it is called an *iso-electronic center*. This recombination center can greatly enhance the probability of radiative transition in indirect-bandgap semiconductors.

Another interpretation of the N -doped bandgap system can be found using the Heisenberg uncertainty principle. One representation of this equation is

$$\Delta p \Delta x \geq \hbar, \quad (25)$$

where p is the momentum and x is the position. Because the position of a localized state due to any nitrogen atom is well known (each one could in theory be identified), Δx is small. If so, Δp is large. This means that the wave function of the level spreads out in k -space and has a finite value directly above the top of the valence band. An electron dropping from the conduction band to the nitrogen level has a finite probability of appearing directly above the top of the valence band, as shown in Fig. 10a. Thus the indirect bandgap semiconductor looks like a direct bandgap material. Figure 10b shows the quantum efficiency, the number of photons generated per electron-hole pair, versus alloy composition for $\text{GaAs}_{1-y}\text{P}_y$ with and without the isoelectronic impurity nitrogen.⁷ The efficiency without nitrogen drops sharply in the composition range $0.4 < y < 0.5$ because the bandgap changes from direct to indirect at $y = 0.45$. The efficiency with nitrogen is considerably higher for $y > 0.5$ but nevertheless decreases steadily with an increasing y because of the increasing separation between the direct and indirect bandgap (Fig. 9b).

For high-brightness blue LEDs (455–492 nm), II-VI compounds such as ZnSe , III-V nitride compounds such as GaN , and IV-IV compounds such as SiC have been investigated. However, their short lifetimes prevent II-VI-based devices from being commercialized at present and the indirect bandgap of SiC results in low-brightness blue LEDs. The most promising candidates are GaN ($E_g = 3.44$ eV) and related III-V nitride semiconductors such as AlGaInN , which have direct bandgaps ranging from 0.7 eV to 6.2 eV with corresponding wavelengths from 200 nm to 1770 nm.⁸ Although there are no lattice-matched substrates for the growth of GaN , high-quality GaN has been grown on sapphire (Al_2O_3) using a low-temperature grown GaN or AlN as a buffer layer.

Figure 11a shows a DH nitride LED grown on a sapphire substrate. Due to the high resistivity of the sapphire substrate, both the n - and p -type ohmic contacts are formed on the top surface. The blue light originates from the radiative recombination $\text{In}_x\text{Ga}_{1-x}\text{N}$ region, which is sandwiched between two larger bandgap semiconductors—a p -type $\text{Al}_x\text{Ga}_{1-x}\text{N}$ layer and an n -type GaN . The higher bandgap p -type $\text{Al}_x\text{Ga}_{1-x}\text{N}$ confinement layer is used to block effectively the electron injected from n - GaN due to the higher conduction band offset. The multiple quantum well $\text{In}_x\text{Ga}_{1-x}\text{N}/\text{GaN}$ LED is shown in Fig. 11b, which has higher quantum efficiency from higher carrier recombination efficiency.

Visible LEDs can be used for full-color displays, full-color indicators, and lamps with high efficiency and high reliability. Figure 12 shows diagrams of two LED lamps.⁸ An LED lamp contains an LED chip and a plastic lens, which is usually colored to serve as an optical filter and to enhance contrast. The lamp in Fig. 12a uses a conventional diode header. Figure 12b shows a package that is suited for a transparent semiconductor, such as GaP and sapphire, which emits light through all five facets (four sides and the top) of the LED chip.

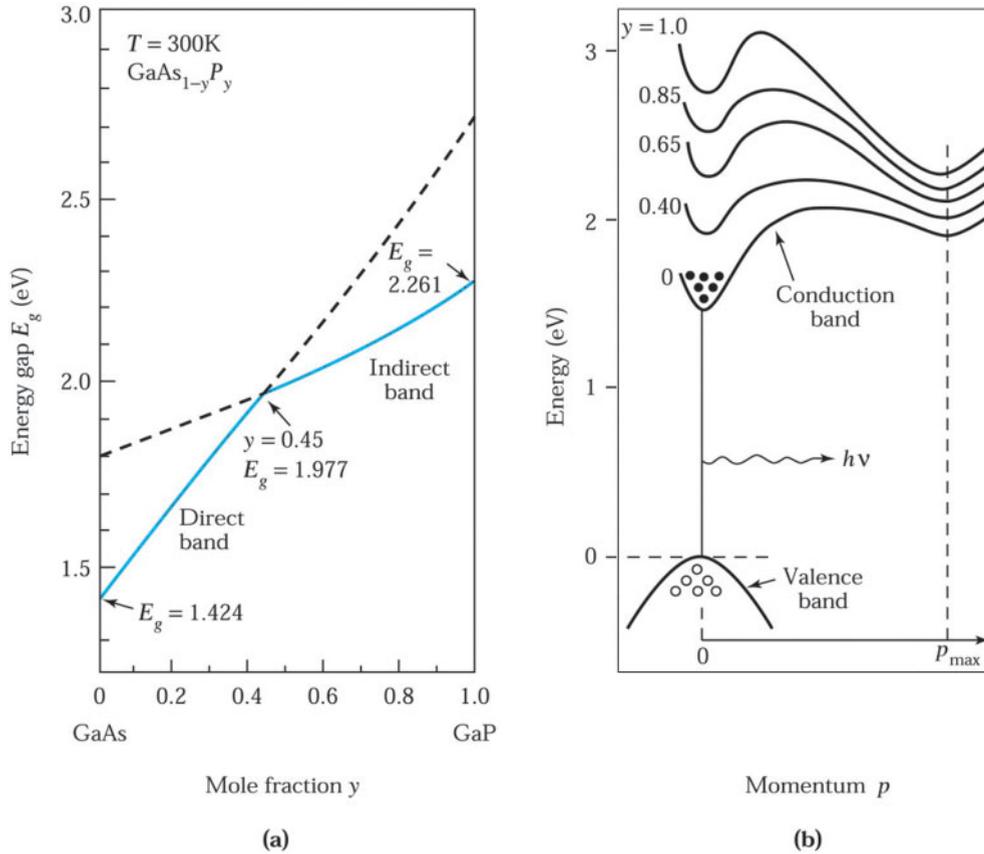


Fig. 9 (a) Compositional dependence for the direct- and indirect-energy bandgap for $\text{GaAs}_{1-y}\text{P}_y$. (b) The alloy compositions shown correspond to red ($y = 0.4$), orange (0.65), yellow (0.85), and green light (1.0).⁶

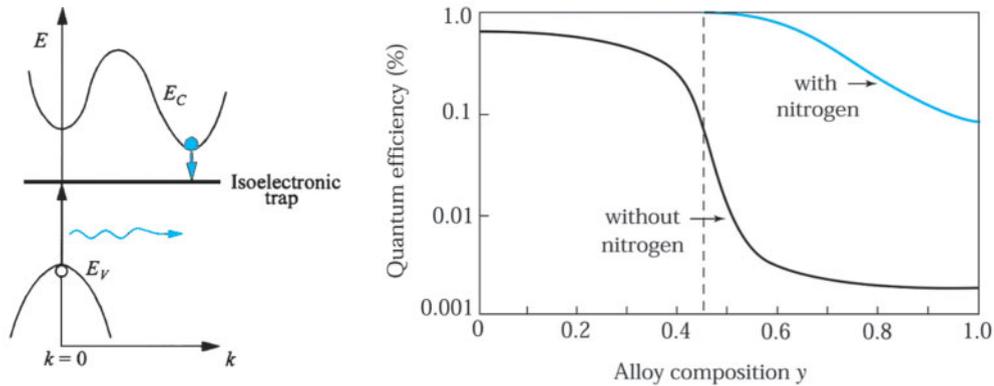


Fig. 10 (a) E - k diagram showing radiative recombination through an isoelectronic trap in indirect-bandgap material. (b) Quantum efficiency versus alloy composition with and without isoelectronic impurity nitrogen.⁷

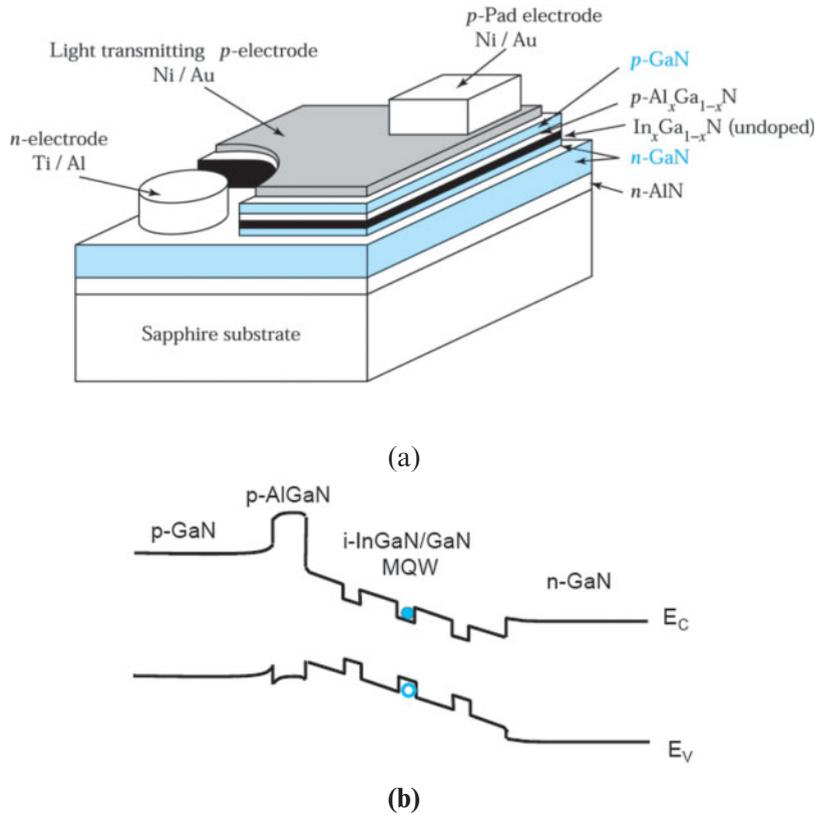


Fig. 11 (a) III-V nitride LED grown on a sapphire substrate. (b) The blue light originates from the multiple quantum well $\text{Ga}_x\text{In}_{1-x}\text{N}/\text{GaN}$ region sandwiched between a p -type $\text{Al}_x\text{Ga}_{1-x}\text{N}$ layer and an n -type GaN layer.

9.3.2 Organic LED

In recent years, certain organic semiconductors have been studied for electroluminescent applications. The organic light-emitting diode (OLED) is particularly useful for a multicolor, large-area flat-panel display because of its attributes of low-power consumption and excellent emissive quality with a wide viewing angle.¹⁰

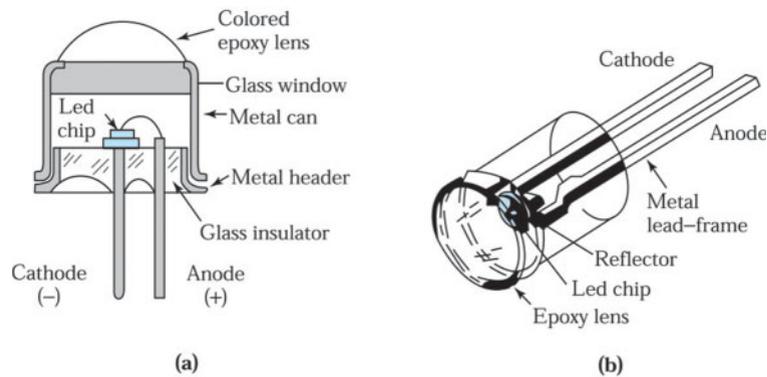


Fig. 12 Diagrams of two LED lamps,⁹ (a) a metal and (b) a plastic package.

OLED and PLED

OLEDs are made from small molecules or polymers. Commonly, macro-molecules with a molecular weight greater than 10,000 atomic mass units (amu) are called polymers, whereas lighter molecules are referred to small molecules. Usually, a polymer light-emitting diode is referred to as a PLED. A small molecule light-emitting diode is referred to as an OLED because the first high-efficiency OLED was made from small molecules. The structures of the OLED prepared by vacuum deposition techniques or the PLED prepared by spin-coating, screen printing etc, are usually amorphous with current preparation methods.

Conductivities of Polymers and Small Molecules

Carbon can form two primary hybrid structures. One structure is the tetrahedrally directed covalent bonds (sp^3 hybridized) where valence electrons are tightly bound and act as insulators, as in diamond and saturated polymers (e.g., ethane, C_2H_6). The other structure is the hexagonally directed covalent bonds (sp^2 hybridized) with planar geometry, as in graphite and conjugated polymers (e.g., ethylene, C_2H_4). The electron orbital will form a weak delocalized π - π bond with neighboring carbon atoms to result in alternating single and double bonds. The structure is said to be conjugated. The π electrons do not belong to a single bond or atom, but rather to a group of atoms. The electrons in the π -bonds are less strongly bound than the electrons in the σ -bonds and have the potential to display either semiconducting or metallic behavior.

Benzene (C_6H_6) has also the molecular structure of sp^2 hybrid orbitals. It is a planar six-carbon ring with alternating single and double bonds, as shown in Fig. 13. The structural of benzene is also conjugated. In OLED, the benzene ring is an important base and is in charge of electron transport within small molecules; however, charge transport across the molecules is ascribed to a hopping process. In most organic semiconductors and unlike inorganic semiconductors, the mobility increases with temperature due to higher thermal energy. For OLED the mobility is low and related to the disordered nature of the solid-state nanostructure. The highest hole mobility is about $15 \text{ cm}^2/\text{V-s}$ and the highest electron mobility is about $0.1 \text{ cm}^2/\text{V-s}$ for a single crystal of small molecules. The molecules packed into well organized polycrystalline films will lead to higher mobility.

Bandgaps

An organic molecule is covered with electrons with a specific spatial distribution and energy, which is the molecular orbital. Electrons occupy molecular orbitals from the lowest first to higher energy level.

HOMO and LUMO are acronyms for highest occupied molecular orbital and lowest unoccupied molecular orbital, respectively. When two molecules interact, a splitting of the HOMO and the LUMO energy levels will be induced. When many molecules interact, a continuum HOMO band and LUMO band corresponding to the valence band and the conduction band of inorganic semiconductors will be formed, similar to inorganic semiconductors discussed in Chapter 1. The energy difference of the highest energy of HOMO band and the lowest energy of LUMO band is the band gap. Various kinds of organic semiconductors have various bandgaps. The optical emission and absorption of an OLED depends on its bandgap.

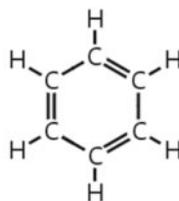


Fig. 13 Structure and delocalized π -bond and σ -bond of benzene.

Only 25% of the transitions are radiative recombinations according to spin statistics, and efficient electroluminescence is difficult to achieve using purely organic materials. An excited molecule (host emitter) can transfer energy to a molecule (guest emitter or dopant) in a lower energy state. If the dopant has higher efficiency, it can enhance the emission efficiency and also change the electroluminescence color and lifetime.

OLED Structures

The high-performance OLED was developed using the concept of multilayer structures. Figure 14a shows the molecular structure of two representative organic semiconductors used for a double-layer structure.¹¹ They are the tris (quinolin-8-olato) aluminum (AlQ₃), and the aromatic diamine. AlQ₃ contains six benzene rings connected to a central aluminum atom, which can strongly attract electrons and creates an electron-deficient state that is an electron transport layer (ETL). The aromatic diamine also contains six benzene rings but with a different molecular arrangement. Nitrogen in the diamine structure has a lone electron pair, which is easily ionized to accept holes. Therefore, diamine is a hole transport layer (HTL). A basic OLED has a number of layers on a transparent substrate (e.g., glass). On this substrate are deposited, in sequence, a transparent conductive anode [e.g., ITO (indium tin oxide)], the diamine as a HTL layer, the AlQ₃ as a ETL layer, and the cathode contact (e.g., Mg alloy with 10% Ag). A cross-sectional view is shown in Fig. 14b. Figure 14c shows the band diagram of the OLED. It is basically a heterojunction formed between AlQ₃ and diamine. Under proper biasing, electrons are injected from the cathode and move toward the heterojunction interface, whereas holes are injected from the anode and also move toward the interface. Because of the energy barriers ΔE_C and ΔE_V , these carriers will accumulate at the interface to enhance the chance of radiative recombination.

The function of the HTL is to assist the injection of holes from the anode, accept these holes, and transport them to the heterojunction interface. Therefore, the energy levels between the HTL and the anode should match for hole injection from the anode and the hole mobility should be high. It is better if the HTL layer also has an electron blocking function. The ETL has the function of assisting the injection of electrons from a metal cathode and transporting them throughout the ETL film. Therefore, the energy levels between the ETL and cathode should match for electrons injection from the cathode and the electron mobility should be high. It is better if the ETL layer also has a hole-blocking function. For the diamine/AlQ₃ bilayer structure, $\Delta E_V < \Delta E_C$. The larger electron barrier ΔE_C effectively blocks the electrons and confines them to the interface. The hole barrier ΔE_V , however, is relatively small and thus can still allow significant amounts of hole injection into the AlQ₃. Therefore, the ETL layer is also an emissive layer (EML). This configuration apparently improves EL efficiency by forcing the recombination to occur in the AlQ₃ and limiting the electron leakage current. It is worth noting that the smaller a hole barrier, the smaller the applied voltage needed for the same current. But the increased hole injection into ETL/EML may not be desired because a larger portion of the holes can leak into the cathode or combine near the cathode where photoquenching centers are abundant. The photoquenching centers, i.e., nonemissive electron-hole recombination centers, are from the carbonyl groups that are formed by an oxidation reaction from the diffusion of oxygen and moisture from air into the AlQ₃ through microscopic pinholes and cracks or grain boundaries in the cathode during the light exposure in the normal ambient.

From Fig. 14c, we can specify the design criteria for an OLED: (a) ultrathin layers for low biasing voltage—for example, the total thickness of the organic semiconductor layers shown is only 150 nm; (b) low injection barriers—the barrier height $q\phi_1$ for hole injection and the barrier height $q\phi_2$ for electron injection must be low enough to allow large carrier injections for high-current density operation, and (c) proper bandgaps for the required color. For AlQ₃, the emitted light is green. By choosing different organic semiconductors with different bandgaps, various colors including red, yellow, and blue can be obtained.

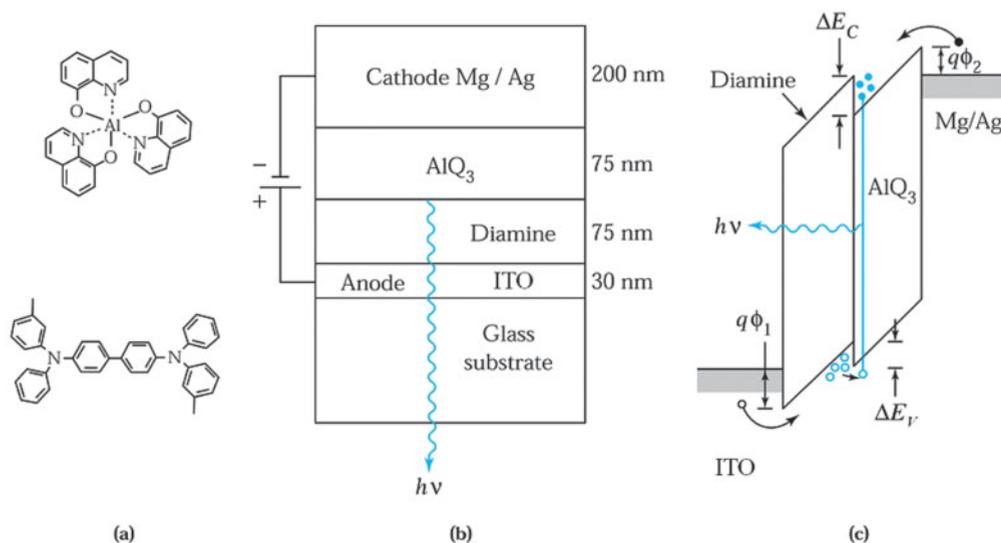


Fig. 14 (a) Organic semiconductors. (b) OLED cross-sectional view. (c) Band diagram of an OLED.

A triple-layer structure may be used where a thin EML is sandwiched between HTL and ETL (ITO/HTL/EML/ETL/Metal), as shown in Fig. 15a. The electron and hole concentrations can be higher in EML and hence the light emission efficiency is higher, as shown in the band diagram of Fig. 15a. The insertion of a thin hole injection layer (HIL) between the ITO and the HTL to lower the barrier height $q\phi_1$ not only lowers the drive voltage but also improves device durability. This is a four-layer structured OLED. The insertion of a thin electron injection layer (EIL) between the metal cathode and the ETL yields a five-layer (ITO/HIL/HTL/EML/ETL/EIL/Metal) OLED with the corresponding band diagram shown in Fig. 15b. This structure will have higher efficiency.

9.3.3 White-Light LED

There has been interest in the development of white LEDs for general illumination because LEDs have much higher efficiency than incandescent lamps. In addition, LEDs can last 10 times longer.

White light can be produced by mixing two or three colors of an appropriate intensity ratio. There are basically two approaches to achieving white light. The first is to combine LEDs of different colors: red, green, and blue. This is not a popular approach since it is more costly and involves sophisticated electro-optical design to control the blending of different colors. The second approach, most commonly used, is to have a single LED covered with a color converter. A color converter is a material that absorbs the original LED light and emits light of different frequency. The converter material can be phosphor, organic dye, or another semiconductor, with phosphor the most common of the three.¹² The light output from a phosphor generally has a much broader spectrum than the LED light. The efficiencies of these color converters can be very high, near 100%. One popular version is to use a blue LED together with a yellow phosphor. In this scheme, the LED light is partially absorbed by phosphor. The blue LED light is mixed with yellow light produced by the phosphor to give white light. Another version is to use a UV LED to stimulate red, green, and blue phosphors to give a white light.

9.3.4 Infrared LED

Infrared LEDs include gallium arsenide LEDs, which emit light near $0.9 \mu\text{m}$, and many III-V compounds, such as the quaternary $\text{Ga}_x\text{In}_{1-x}\text{As}_y\text{P}_{1-y}$ LEDs, which emit light from 1.1 to $1.6 \mu\text{m}$.

An important application of infrared LEDs is in opto-isolators, where an input or control signal is decoupled from the output. Figure 16 shows an opto-isolator having an infrared LED as the light source and a photodiode as the detector. When an input signal is applied to the LED, light is generated and subsequently detected by the

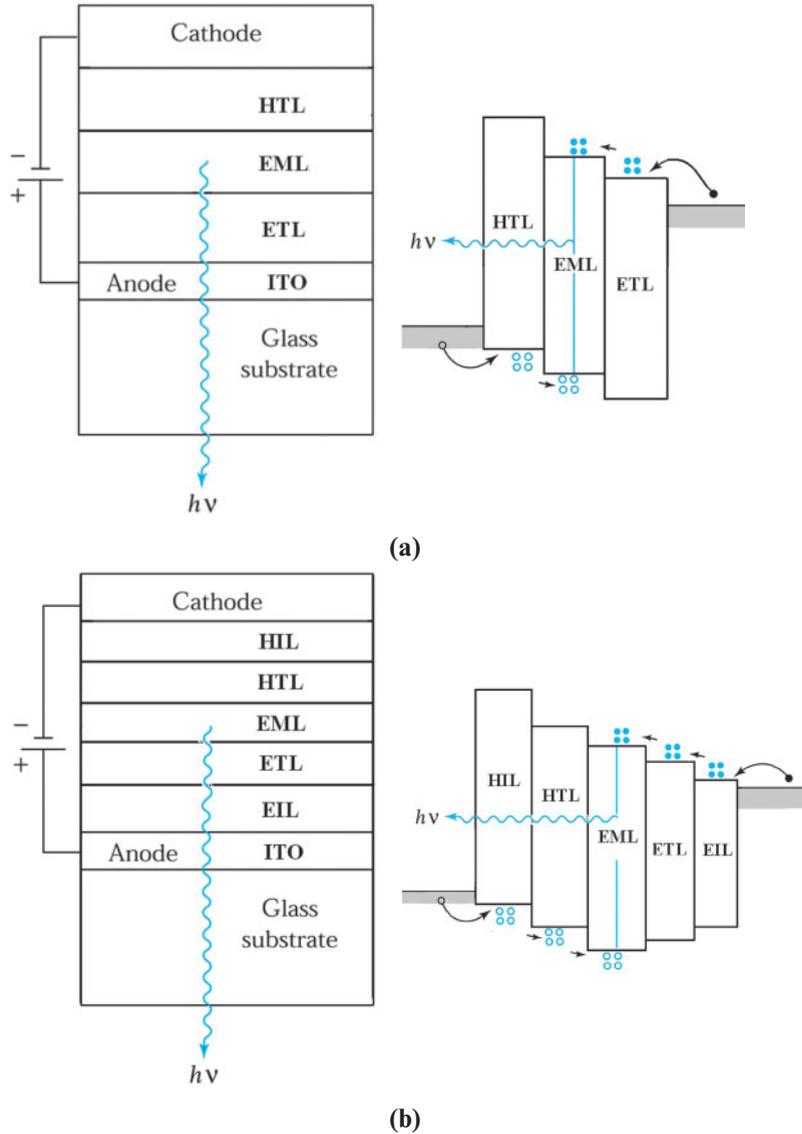


Fig. 15 OLEDs with (a) three-layer and (b) five-layer structures.

photodiode. The light is then converted back to an electrical signal as a current that flows through a load resistor. Opto-isolators transmit signals at the speed of light and are electrically isolated because there is no electrical feedback from the output to the input.

Another important application of infrared LEDs is for transmission of an optical signal through an optical fiber, as in a communication system. An optical fiber is a waveguide at optical frequencies. The fiber is usually drawn from a preform of glass to a diameter of about $100\ \mu\text{m}$. It is flexible and can guide optical signals over distances of many kilometers to a receiver, similarly to the way a coaxial cable transmits electrical signals.

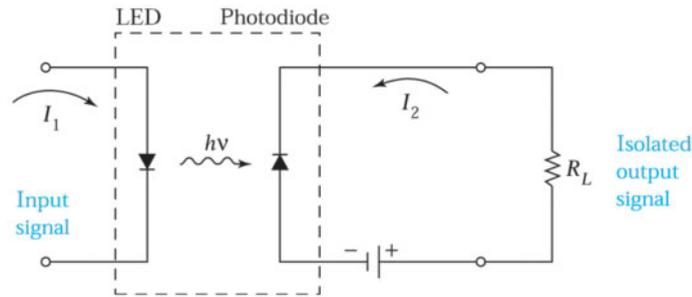


Fig. 16 An opto-isolator in which an input signal is decoupled from the output signal.

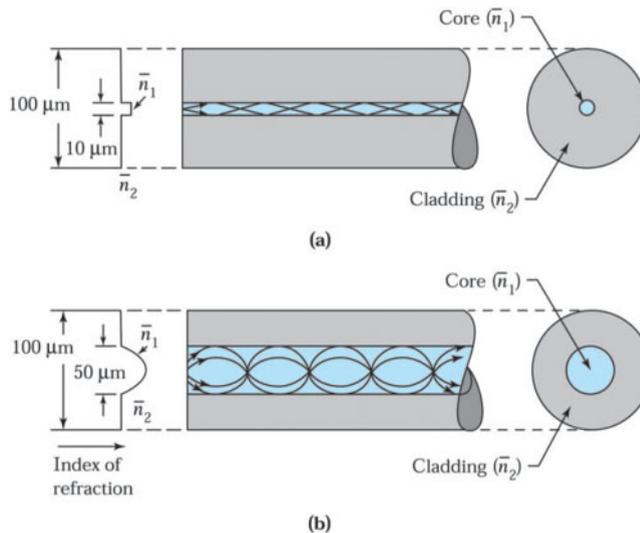


Fig. 17 Optical fibers. (a) Step-index fiber having a core with slightly larger reflective index. (b) Graded-index fiber having a parabolic grading of the reflective index in the core.¹³

Two types of optical fibers are shown in Fig. 17. One type of fiber has a cladding layer of relatively pure fused silica (SiO_2) surrounding a core of doped glass (e.g., germanium doped glass) with a higher refractive index than the cladding layer.¹³ This type of fiber is called a step-index fiber. The light is transmitted along the length of the fiber by internal reflection at the step in the refractive index. The critical angle for internal reflection is about 79° for $\bar{n}_1 = 1.457$ (cladding layer) and $\bar{n}_2 = 1.480$ (core, 20% Ge-doped), as calculated from Eq. 23. Note that different rays will propagate with different path lengths (Fig. 17a). A light pulse reaching the end of a step index fiber will result in a pulse spread. In a graded-index fiber (Fig. 17b), the index decreases from the core center by a parabolic law. Now, rays traversing toward the cladding have a higher velocity (due to lower refractive index) than rays along the center of the core. The pulse spread is significantly reduced. As the light is transmitted along the optical fiber, the light signal will be attenuated. However, due to the transparency of ultrapure silica used for the fiber material in the wavelength region from 0.8 to 1.6 μm , the attenuation is quite low and is proportional to λ^{-4} . Typical attenuations are about 3 dB/km at a wavelength of 0.8 μm , 0.6 dB/km at 1.3 μm , and 0.2 dB/km at 1.55 μm .¹⁴

A simple point-to-point optical-fiber communication system is shown in Fig. 18, where the electrical input signals are converted to optical signals using an optical source (LED or laser). The optical signals are coupled into the fiber and transmitted to the photodetector, where they are converted back to electrical signals.

The surface-emitting infrared InGaAsP LED used for optical-fiber communication is shown¹⁵ in Fig. 19. The light is emitted from the central surface area and coupled into the optical fiber. The use of heterojunctions (e.g., InGaAsP-InP) can increase the efficiency that results from the confinement of the carrier by the layers of the higher-bandgap semiconductor InP surrounding the radiative-recombination region InGaAsP. The heterojunction can also serve as an optical window to the emitted radiation because the higher-bandgap-confining layers do not absorb radiation from the lower-bandgap-emitting region.

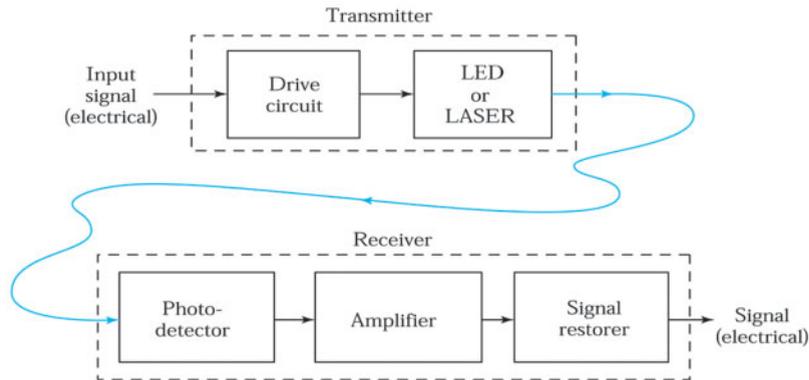


Fig. 18 Basic elements of an optical fiber transmission link.

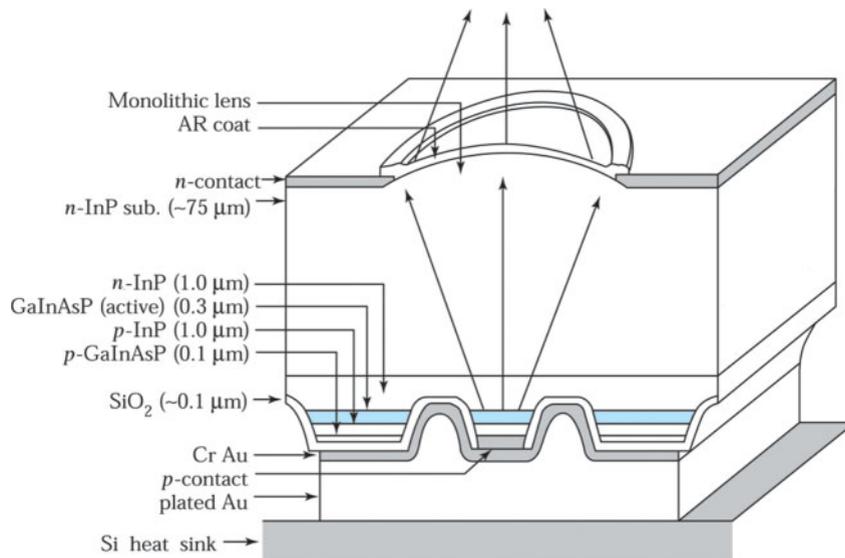


Fig. 19 Small-area mesa-etched GaInAsP/InP surface-emitting LED structure.¹⁵

► 9.4 SEMICONDUCTOR LASERS

Semiconductor lasers are similar to the solid-state ruby laser and helium-neon gas laser in that the emitted radiation is highly monochromatic and produces a highly directional beam of light. However, the semiconductor laser differs from other lasers in that it is small (on the order of 0.1 mm long) and is easily modulated at high frequencies simply by modulating the biasing current. Because of these unique properties, the semiconductor laser is one of the most important light sources for optical-fiber communication. It is also used in video recording, optical reading, and high-speed laser printing. In addition, semiconductor lasers have significant applications in many areas of basic research and technology, such as high-resolution gas spectroscopy and atmospheric pollution monitoring.

9.4.1 Semiconductor Materials

All lasing semiconductors have direct bandgaps. This is expected because the momentum is conserved and the radiative-transition probability in a direct-bandgap semiconductor is high. At present, the laser emission wavelengths cover the range from 0.3 to over 30 μm . Gallium arsenide was the first material to emit laser radiation and its related III-V compound alloys are the most extensively studied and developed.

The three most important III-V compound alloy systems are $\text{Ga}_x\text{In}_{1-x}\text{As}_y\text{P}_{1-y}$, $\text{Ga}_x\text{In}_{1-x}\text{As}_y\text{Sb}_{1-y}$ and $\text{Al}_x\text{Ga}_{1-x}\text{As}_y\text{Sb}_{1-y}$ solid solutions. Figure 20 shows the bandgaps plotted against the lattice constant for the three alloy systems and their binary, ternary and quaternary compounds.¹⁶ To achieve a heterostructure with negligible interface traps, the lattices between the two semiconductors must be matched closely.

If we use GaAs ($a = 5.6533 \text{ \AA}$) as the substrate, the ternary compound $\text{Al}_x\text{Ga}_{1-x}\text{As}$ can have a lattice mismatch less than 0.1% for $0 \leq x \leq 1$. With InP ($a = 5.8687 \text{ \AA}$) as the substrate, the quaternary compound $\text{Ga}_x\text{In}_{1-x}\text{As}_y\text{P}_{1-y}$ can have a nearly perfect lattice match, as indicated by the center vertical line in Fig. 20.

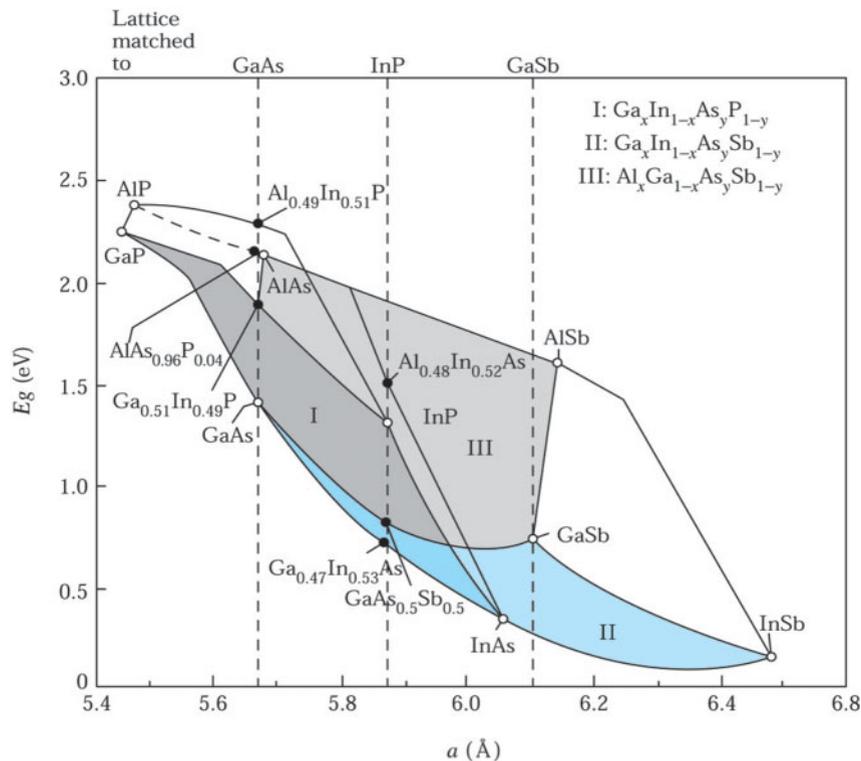


Fig. 20 Energy bandgap and lattice constant for three III-V compound solid alloy system.¹⁶

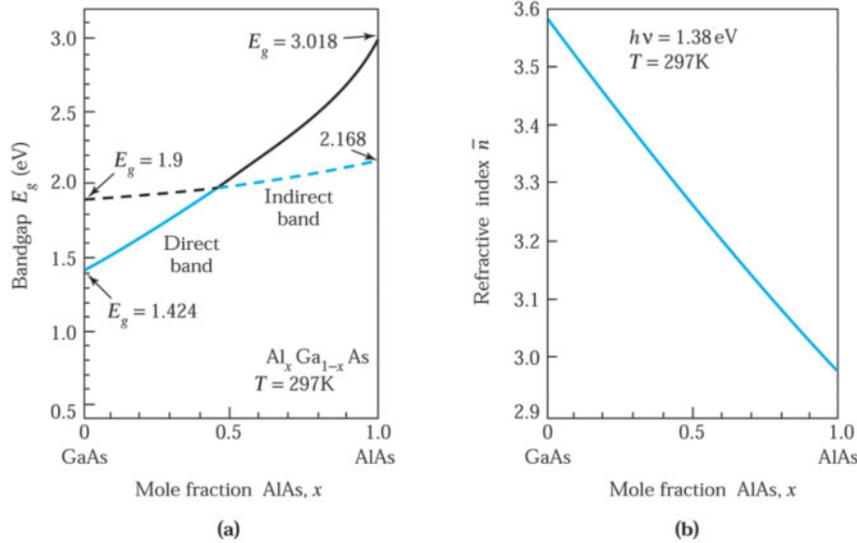


Fig. 21 (a) Compositional dependence of the $\text{Al}_x\text{Ga}_{1-x}\text{As}$ energy gap.¹ (b) Compositional dependence of the refractive index at 1.38 eV.

Figure 21a shows the bandgap of ternary $\text{Al}_x\text{Ga}_{1-x}\text{As}$ as a function of aluminum composition.¹ The alloy has a direct bandgap up to $x = 0.45$, then becomes an indirect-bandgap semiconductor. Figure 21b shows the compositional dependence of the refractive index. Basically, the refractive index is inversely proportional to the bandgap. For example, for $x = 0.3$, the bandgap of $\text{Al}_x\text{Ga}_{1-x}\text{As}$ is 1.789 eV, which is 0.365 eV larger than that of GaAs; its refractive index is 3.385, which is 6% smaller than that of GaAs. These properties are important for continuous operation of semiconductor lasers at and above room temperature.

The new class of nitride-based materials (AlGaIn and AlInN) made a significant advance in the past decade. Blue lasers usually operate at 405 nm but in general the operation of these devices is demonstrated between 360 and 480 nm. These devices have applications in many areas ranging from optical data storage in high-density digital video disc (HD DVD) to medical applications. Infrared and red lasers with a wavelength between 780 nm and 650 nm are currently used for optical data storage. To increase the capacity of optical disks, laser diodes with much shorter wavelengths are needed because the minimum spot size of focused light is limited by the wavelength of the light.

9.4.2 Laser Operation

Figure 22 shows schematic representations¹⁷ of the band diagram under forward-bias, the refractive index profile, and the optical-field distribution of light generated at the junction of a homojunction laser (Fig. 22a) and a double-heterostructure (DH) laser (Fig. 22b).

Population Inversion

As discussed in Section 9.1.1, to enhance the stimulated emission for laser operation we need population inversion. To achieve population inversion in a semiconductor laser, we consider a p - n junction or a double heterojunction (DH) formed between degenerate semiconductors. This means that the doping levels on both sides of the junction are high enough that the Fermi level E_{FV} is below the valence band edge on the p -side and E_{FC} is above the conduction band edge on the n -side. When a sufficiently large bias is applied (the band diagrams in Fig. 22), high injection occurs: that is, large concentrations of electrons and holes are injected into the transition region. As a result, the region d (Fig. 22) contains a large concentration of electrons in the conduction band and a large concentration of holes in the valence band; this is the required condition for population inversion. For band-to-band transition, the minimum energy required is the bandgap energy E_g . Therefore, from the band diagram in Fig. 22, we can write the condition necessary for population inversion: $(E_{FC} - E_{FV}) > E_g$.

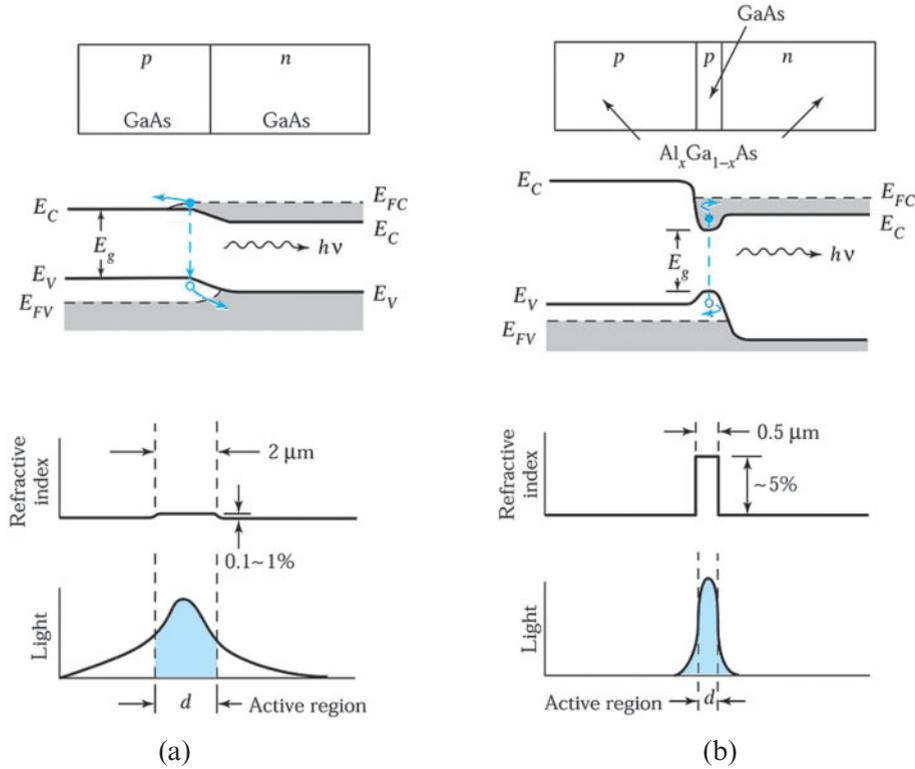


Fig. 22 Comparison of some characteristics of (a) homojunction laser and (b) double-heterostructure (DH) laser. The second from the top row shows energy band diagrams under forward bias. The refractive index change for a homojunction laser is less than 1%. The refractive index change for DH laser is about 5%. The confinement of light is shown in the bottom row.¹⁷

Carrier and Optical Confinement

As can be seen in the DH laser, the carriers are confined on both sides of the active region by the heterojunction barriers, whereas in the homojunction laser the carriers can move away from the active region, where radiative recombination occurs.

For the homojunction laser, the difference in the refractive indices between the center waveguiding layer and the adjacent layers arises from the difference of carrier density. Material with higher carrier density has a lower refractive index. Here the carrier density of the active layer is less than that heavily doped n^+ - and p^+ -layers. The refractive index change for a homojunction laser is only 0.1% to about 1%. In the DH laser the optical field is confined within the active region by the abrupt reduction of the refractive index outside the active region. The optical confinement can be explained by Fig. 23, which shows a three-layer dielectric wave guide with refractive indices \bar{n}_1 , \bar{n}_2 and \bar{n}_3 , where an active layer is sandwiched between two confining layers (Fig. 23a). Under the condition $\bar{n}_2 > \bar{n}_1 \geq \bar{n}_3$, the ray angle θ_{12} at the layer 1/layer 2 interface in Fig. 23b exceeds the critical angle given by Eq. 23. A similar situation occurs for θ_{23} at the layer 2/ layer 3 interface. Therefore, when the refractive index in the active layer is larger than the index of its surrounding layers, the propagation of the optical radiation is guided (confined) in a direction parallel to the layer interfaces. We can define a *confinement factor* Γ as the ratio of the light intensity within the active layer to the sum of light intensity both within and outside the active layer. The confinement factor is given as

$$\Gamma \cong 1 - \exp(-C\Delta\bar{n}d), \tag{26}$$

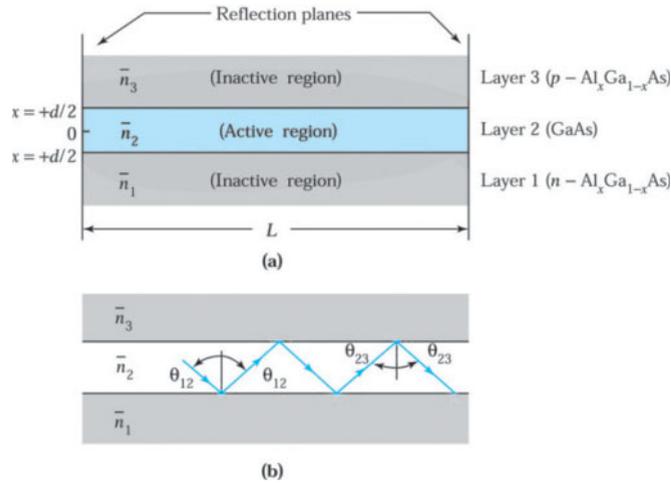


Fig. 23 (a) Representation of a three-layer dielectric waveguide. (b) Ray trajectories of the guided wave.

where C is a constant, $\Delta\bar{n}$ is the difference in the refractive index, and d is the thickness of the active layer. It is clear that the larger $\Delta\bar{n}$ and d are, the higher Γ will be.

Optical Cavity and Feedback

We have considered the condition necessary to produce laser action: population inversion. Photons released by stimulated emission are likely to cause further stimulations as long as there is population inversion. This is the phenomenon of optical gain. The gain obtained in a single travel of an optical wave down a laser cavity is small. To increase gain, multiple passes of a wave must occur. This is achieved using mirrors placed at either end of the cavity, shown as the reflection planes at the left side and right side in Fig. 23a. For a semiconductor laser, the cleaved ends of the crystal forming the device can act as the mirrors. For a GaAs device, cleaving along (110) plane creates two parallel identical mirrors. Sometimes the back mirror of the laser is metallized to enhance the reflectivity. The reflectivity R at each mirror can be calculated as

$$R = \left(\frac{\bar{n} - 1}{\bar{n} + 1} \right)^2, \quad (27)$$

where \bar{n} is the refractive index in the semiconductor corresponding to the wavelength λ (\bar{n} is generally a function of λ).

► EXAMPLE 3

Calculate the R for GaAs ($\bar{n} = 3.6$).

SOLUTION From Eq. 27,

$$R = \left(\frac{3.6 - 1}{3.6 + 1} \right)^2 = 0.32,$$

that is, 32% of the light will be reflected at the cleaved surface. ◀

If an integral number of half-wavelengths fits between the two end planes, reinforced and coherent light will be reflected back and forth within the cavity. Therefore, for stimulated emission, the length L of the cavity must satisfy the condition

$$m \left(\frac{\lambda}{2\bar{n}} \right) = L \tag{28}$$

or

$$m\lambda = 2\bar{n}L, \tag{28a}$$

where m is an integral number. Obviously, many values of λ can satisfy this condition (Fig. 24a), but only those within the spontaneous emission spectrum will be produced (Fig. 24b). In addition, optical losses in the path traveled by the wave mean that only the strongest lines will survive, leading to a set of lasing modes, as shown in Fig. 24c. These modes are called longitudinal modes, as they occur because of standing waves formed in the longitudinal direction of a laser diode. The separation $\Delta\lambda$ between the allowed modes in the longitudinal direction is the difference in the wavelengths corresponding to m and $m + 1$. Differentiating Eq. 28a with respect to λ , we obtain

$$\Delta\lambda \cong \frac{\lambda^2 \Delta m}{2\bar{n}L [1 - (\lambda/\bar{n})(d\bar{n}/d\lambda)]}. \tag{29}$$

Although \bar{n} is a function of λ , over the very small change in wavelength between adjacent modes $d\bar{n}/d\lambda$ is very small, and hence to a good approximation the mode spacing $\Delta\lambda$ is given by (for $\Delta m = 1$)

$$|\Delta\lambda| \cong \frac{\lambda^2}{2\bar{n}L}. \tag{30}$$

As a typical laser is operated at low currents, the spontaneous emission has broad spectral distribution with a full width of half-maximum intensity of 5 to 20 nm. It is similar to emission in an LED. As the bias current approaches the threshold value, the optical gain can be high enough for amplification so that intensity peaks start to appear. At this bias level, the light is still incoherent due to the nature of spontaneous emission. When the bias reaches the threshold current, the lasing spectra suddenly become much narrower ($< 1 \text{ \AA}$), as shown in Fig. 24c, and the light is coherent and much more directional. The number of longitudinal modes can be reduced with further increase of bias current.

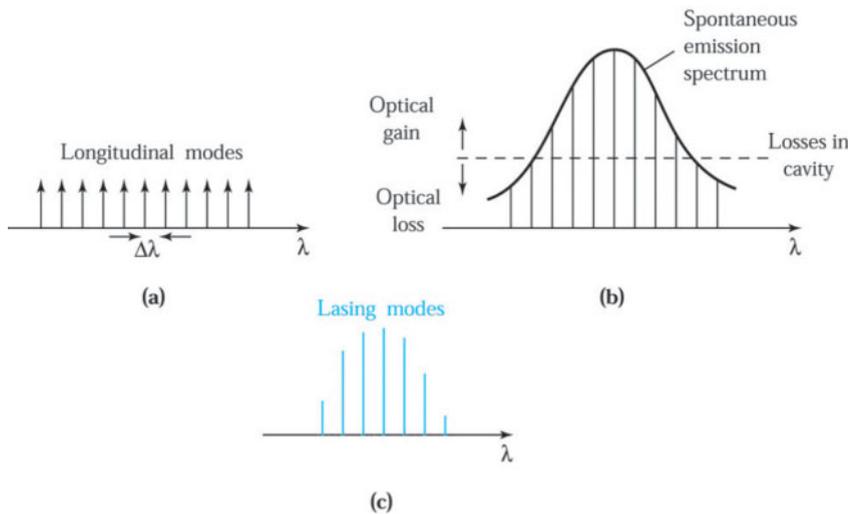


Fig. 24 (a) Resonant modes of a laser cavity. (b) Spontaneous emission spectrum. (c) Optical-gain wavelengths.

▶ **EXAMPLE 4**

Calculate the mode spacing for a typical GaAs laser with $\lambda = 0.94 \mu\text{m}$, $\bar{n} = 3.6$, and $L = 300 \mu\text{m}$.

SOLUTION From Eq. 30,

$$\Delta\lambda \cong \frac{(0.94 \times 10^{-6})^2}{2 \times 3.6 \times 300 \times 10^{-6}} = 4 \times 10^{-10} \text{ m} = 4 \text{ \AA}.$$

9.4.3 Basic Laser Structure

Figure 25 shows three laser structures.^{17,18} The first structure (Fig. 25a) is a basic *p-n* junction laser and is called a homojunction laser because it has the same semiconductor material (e.g., GaAs) on both sides of the junction. Under appropriate biasing conditions laser light will be emitted from these planes (only the front emission is shown in Fig. 25). The two remaining sides of the diode are roughened to eliminate lasing in the directions other than the main ones. This structure is called a *Fabry-Perot cavity*, with a typical cavity length L of about $300 \mu\text{m}$. The Fabry-Perot cavity configuration is used extensively for modern semiconductor lasers.

Figure 25b shows a double-heterostructure (DH) laser, in which a thin layer of a semiconductor (e.g., GaAs) is sandwiched between layers of a different semiconductor (e.g., $\text{Al}_x\text{Ga}_{1-x}\text{As}$). The laser structures shown in Figs. 25a and b are broad-area lasers because the entire area along the junction plane can emit radiation. Figure 25c shows a DH laser with a stripe geometry. The oxide layer isolates all but the stripe contact; consequently the lasing area is restricted to a narrow region under the contact. The stripe widths S are typically $5\text{--}30 \mu\text{m}$. The advantages of the stripe geometry are reduced operating current, elimination of multiple-emission areas along the junction, and improved reliability that is the result of removing most of the junction perimeter. Due to the narrow active region, there is a substantial diffraction of the output at the interface with the air and the light output becomes a broad beam.

Threshold Current Density

One of the most important parameters for laser operation is the threshold current density J_{th} , which is the minimum current density required for lasing to occur. Figure 26 shows J_{th} versus operating temperature for a homojunction laser and a DH laser.¹⁷ Note that as the temperature increases, J_{th} for the DH laser increases much more slowly than J_{th} for the homojunction laser. Because of the low values of J_{th} for DH lasers at 300 K, DH lasers can be operated continuously at room temperature. This characteristic has led to the increased use of semiconductor lasers, especially in optical-fiber communication systems.

In a semiconductor laser, the gain g , the incremental optical energy flux per unit length, depends on the current density. The gain g can be expressed as a function of a nominal current density J_{nom} , which is defined for unity quantum efficiency (i.e., number of carriers generated per photon, $\eta = 1$) as the current density required to uniformly excite a $1 \mu\text{m}$ thick active layer. The actual current density is then given by

$$J(\text{A/cm}^2) = \frac{J_{nom}d}{\eta}, \quad (31)$$

where d is the thickness of the active layer in μm . Figure 27 shows the calculated gain for a typical gallium arsenide DH laser.¹⁹ The gain increases linearly with J_{nom} for $50 < g < 400 \text{ cm}^{-1}$. The linear dashed line can be written as

$$g = (g_0/J_0)(J_{nom} - J_0), \quad (32)$$

where $g_0/J_0 = 5 \times 10^{-2} \text{ cm}\cdot\mu\text{m}/\text{A}$ and $J_0 = 4.5 \times 10^3 \text{ A/cm}\cdot\mu\text{m}$.

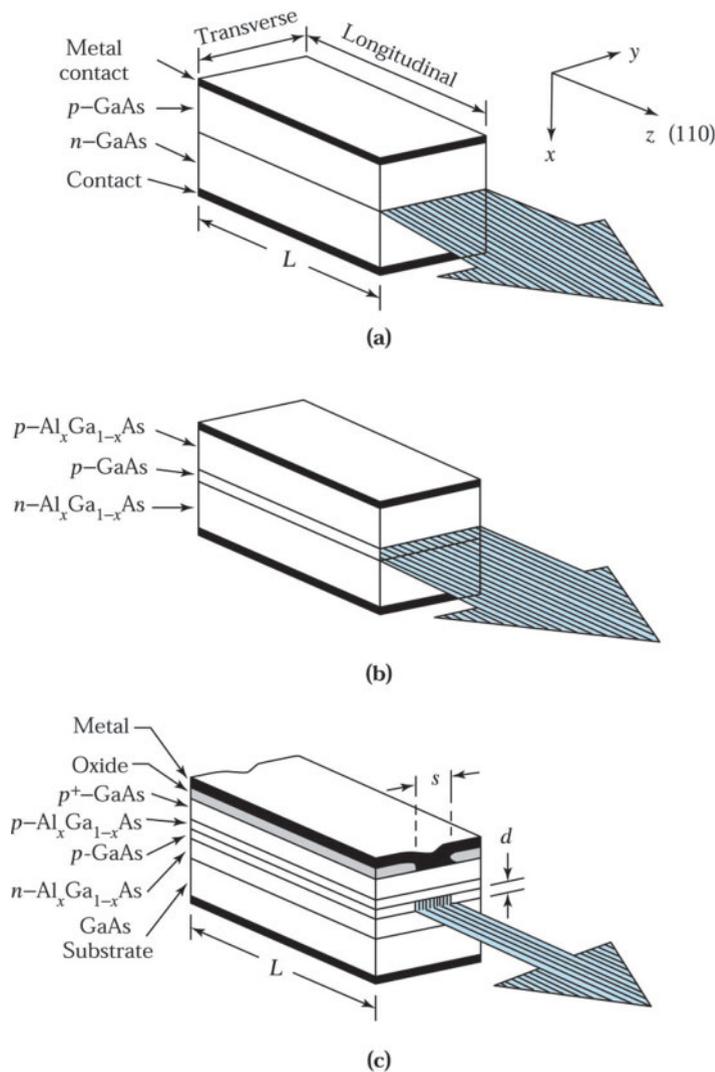


Fig. 25 Semiconductor laser structure in the Fabry-Perot-cavity configuration. (a) Homojunction laser. (b) Double-heterojunction (DH) laser. (c) Stripe-geometry DH laser.^{17,18}

As discussed previously, at low currents we have spontaneous emissions in all directions. As the current density increases, the gain increases (Fig. 27) until the threshold for lasing is reached, that is, until the gain satisfies the condition that a light wave makes a complete traversal of the cavity without attenuation:

$$R \exp[(\Gamma g - \alpha)L] = 1 \tag{33}$$

or

$$\Gamma g(\text{threshold gain}) = \alpha + \frac{1}{L} \ln\left(\frac{1}{R}\right), \tag{34}$$

where Γ is the confinement factor, α is the loss per unit length from absorption and other scattering mechanisms, L is the length of the cavity shown in Fig. 25, and R is the reflectance of the ends of the cavity

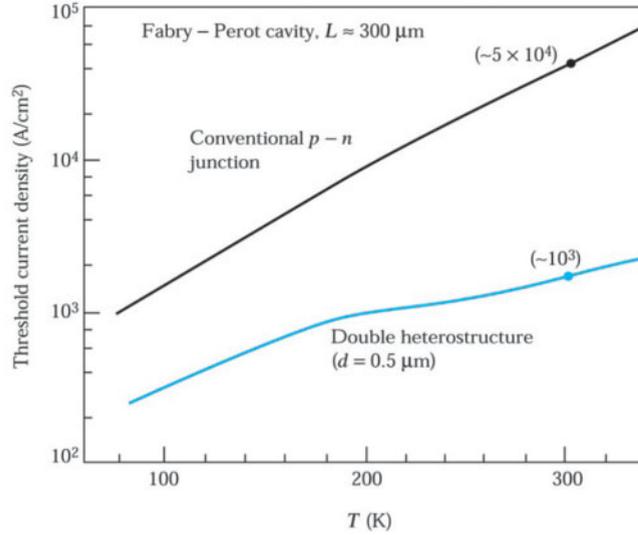


Fig. 26 Threshold current density versus temperature for the two laser structures shown¹⁷ in Fig. 25.

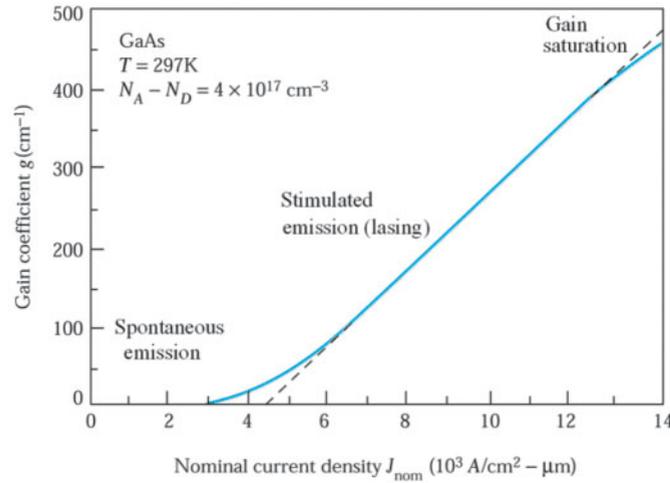


Fig. 27 Variation of gain coefficient versus nominal current density. Dashed line represents linear dependence.¹⁹

assuming that R for both ends is equal. Equations 31, 32 and 34 may be combined to give the threshold current density as

$$J_{th} \text{ (A / cm}^2\text{)} = \frac{J_0 d}{\eta} + \left(\frac{J_0 d}{g_0 \eta \Gamma} \right) \left[\alpha + \frac{1}{L} \ln \left(\frac{1}{R} \right) \right]. \quad (35)$$

The term $(J_0 d / g_0 \eta \Gamma)$ is often called $1/\beta$, where β is known as the gain factor. To reduce J_{th} , we can increase η , Γ , L , and R and reduce d and α .

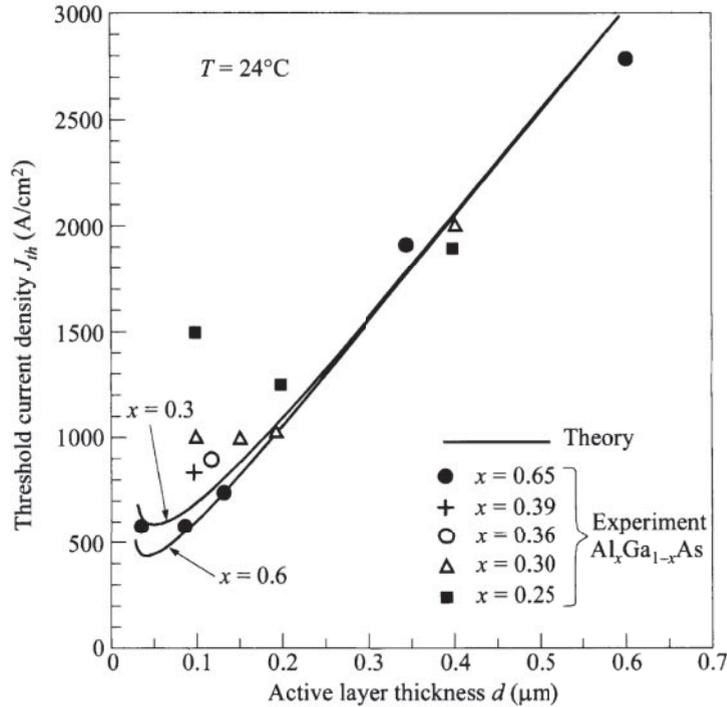


Fig. 28 Comparison of experimental and calculated threshold current density.¹

Figure 28 compares the calculated J_{th} from Eq. 35 to experimental results from $\text{Al}_x\text{Ga}_{1-x}\text{As}$ -GaAs DH lasers.¹ The J_{th} decreases with decreasing d , reaching a minimum, and then increases again. The increase of J_{th} at very small active layer thickness is caused by the poor confinement factor Γ . For a given d , J_{th} decreases with increasing Al composition x because of the improved optical confinement.

► EXAMPLE 5

Find the threshold current for a laser diode using the following data: front and mirror reflectivities are 0.44 and 0.99, respectively. The cavity length and width are 300 μm and 5 μm , respectively, $\alpha = 100 \text{ cm}^{-1}$, $\beta = 0.1 \text{ cm}^{-3}\text{A}^{-1}$, $g_0 = 100 \text{ cm}^{-1}$, and $\Gamma = 0.9$.

SOLUTION With a known gain factor, the term $J_0 d / \eta$ in Eq. 35 can be expressed as $g_0 \Gamma / \beta$. Due to different reflectivities of the two mirrors, Eq. 35 is modified to

$$J_{th} (\text{A} / \text{cm}^2) = \frac{g_0 \Gamma}{\beta} + \frac{1}{\beta} \left[\alpha + \frac{1}{2L} \ln \left(\frac{1}{R_1 R_2} \right) \right] \quad (35a)$$

$$\text{Thus, } J_{th} = \frac{100 \times 0.9}{0.1} + 10 \times \left[100 + \frac{1}{2 \times 300 \times 10^{-4}} \ln \left(\frac{1}{0.44 \times 0.99} \right) \right] = 2036 \text{ A} / \text{cm}^2,$$

$$\text{and so } I_{th} = 2036 \times 300 \times 10^{-4} \times 5 \times 10^{-4} = 30 \text{ mA.} \quad \blacktriangleleft$$

Temperature Effect

Figure 29 shows the temperature dependence of the threshold current I_{th} for a cw (continuous wave) stripe

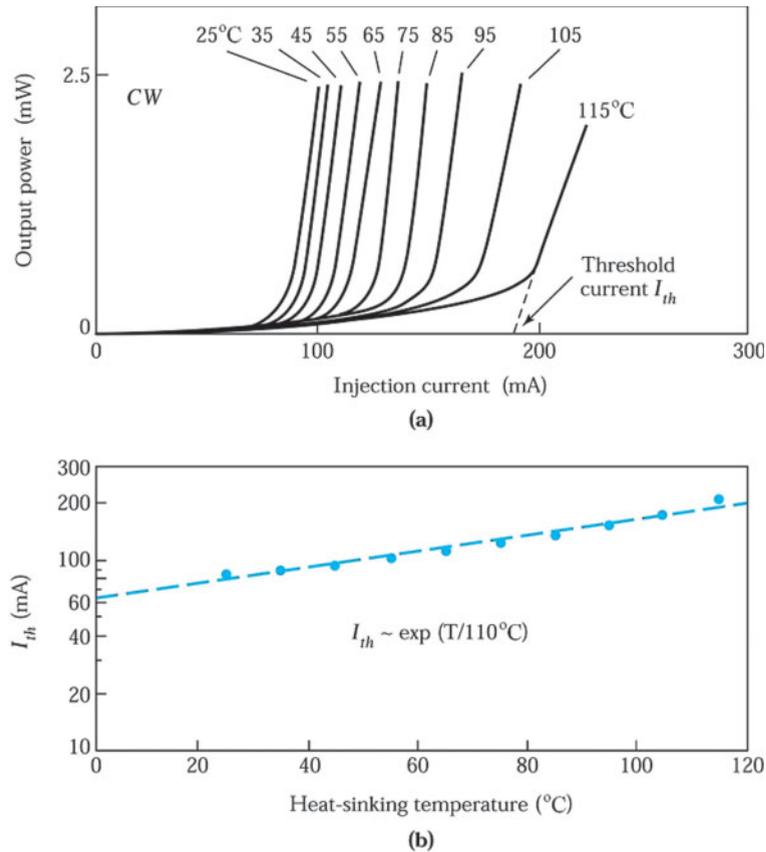


Fig. 29 (a) Light output versus diode current for a GaAs/AlGaAs heterostructure laser. (b) Temperature dependence of the continuous wave (cw) current threshold.²⁰

geometry Al_xGa_{1-x}As-GaAs DH laser.²⁰ Figure 29a shows cw light outputs versus injection current at various temperatures between 25° and 115°C. Note the excellent linearity in the light-current characteristics. The threshold current at a given temperature is the extrapolated value for zero output power. Figure 29b shows a plot of threshold currents as a function of temperature. The threshold current increases exponentially with temperature as

$$I_{th} \sim \exp\left(\frac{T}{T_0}\right), \quad (36)$$

where T is the temperature in °C and T_0 is 110°C for this laser.

► EXAMPLE 6

Calculate the temperature at which the room-temperature value of the threshold current doubles for the laser shown in Fig. 29.

SOLUTION

$$\frac{J_{th}}{2J_{th}} = \frac{\exp(27/110)}{\exp(T/110)}$$

Therefore $T = 27 + 110 \times \ln 2 = 27 + 76 = 103$ °C.

Modulation Frequency and Longitudinal Modes

For optical fiber communications, the optical source must be modulated at high frequencies. Unlike LEDs, whose output power decreases with increasing modulation bandwidth (Eq. 13), the output power of typical GaAs or GaInAsP laser remains at a constant level (e.g., 10 mW per facet) well into GHz range.

For a stripe-geometry GaInAs-AlGaAs DH laser at a current above the threshold, many emission lines exist that are approximately evenly spaced with a separation of $\Delta\lambda$ (e.g., $\Delta\lambda = 4 \text{ \AA}$ in Ex. 4). These emission lines belong to the longitudinal modes given in Eq. 29. Because of these longitudinal modes, the stripe geometry laser is not a spectrally pure light source. For optical-fiber communication systems, an ideal light source is one that has a single frequency. This is because light pulses of different frequencies travel through optical fiber at different speeds, thus causing pulse spread.

9.4.4 Distributed Feedback Lasers

Because of the multimodes in stripe-geometry lasers, these devices are useful only for telecommunication systems operated at relatively low rates (i.e., below 1 Gbit/s). For advanced optical fiber systems, single-frequency lasers are necessary. A single frequency laser operates in only one longitudinal mode. The fundamental approach is to take a laser cavity that allows only one mode to resonate and to provide a constructive interference mechanism that picks out a single frequency. Two laser configurations use this approach—the distributed Bragg reflector (DBR) laser and the distributed feedback (DFB) laser, as shown in Fig. 30.²¹

The DBR is a mirror that has been designed like a reflection type diffraction grating, which has a period corrugated structure. The diffraction grating is somewhat like the double-slit arrangement, but has a much greater number of slits. When monochromatic light is sent through the slits, it forms narrow interference fringes. Diffraction gratings can also be opaque surfaces with narrow parallel grooves arranged like the slits. Light then scatters back from the grooves to form interference fringes rather than being transmitted through open slits. Such reflectors act as frequency-selective mirrors because the constructive and destructive diffraction interference patterns are extremely sensitive to the wavelength of light. The particular Fabry-Perot cavity mode close to λ_B can lase and exist in the output.

Figure 30a shows the cross section of a distributed Bragg reflector (DBR) laser. The region that conducts electric current is called the pumped region. A wavelength-selection grating is placed outside the pumped region. Because of efficient coupling between the active region and the passive grating structure, the reflection is enhanced at the wavelength λ_B , known as the Bragg wavelength, which is related to the period of the grating Λ by

$$\lambda_B = \frac{2\bar{n}\Lambda}{l} \quad (37)$$

where \bar{n} is the effective refractive index of the mode and l is the integer order of the grating. The mode at the Bragg wavelength that has the lowest loss, and thus the lowest threshold gain, will have the dominant output.

Figure 30b shows the distributed feedback (DFB) laser, which has a corrugated grating structure within the active region. The grating region has a periodically varying index of refraction that enhances the wavelength closest to the Bragg wavelength, thus achieving single-frequency operation. Because of the small temperature dependence of the refractive index, the lasing wavelength of the DFB laser has a very small temperature coefficient ($\sim 0.5 \text{ \AA}/^\circ\text{C}$), while the temperature coefficient for a corresponding stripe-geometry laser is much larger ($\sim 3 \text{ \AA}/^\circ\text{C}$), because it follows the temperature dependence of the bandgap. DBR and DFB lasers are also useful as optical sources in integrated optics, which uses miniature optical waveguide components and circuits made by planar technology on rigid substrates.

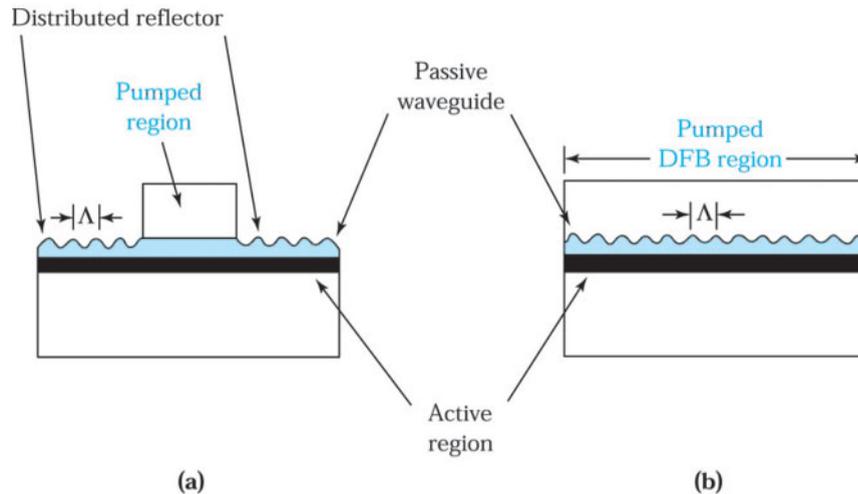


Fig. 30 Two methods of obtaining a single-frequency laser. (a) Distributed Bragg reflector (DBR) laser, and (b) a distributed feedback (DFB) laser.

9.4.5 Quantum-Well Lasers

The structure of a quantum-well (QW) laser^{21,22} is similar to that of a DH laser except that the thickness of the active layer in a QW laser is very small, about 10–20 nm. Figure 31a shows the band diagram of a QW laser where the central GaAs region ($L_y \cong 20$ nm) is sandwiched between two larger bandgap AlGaAs layers. The length L_y is comparable to the de Broglie wavelength ($\lambda = h/p$, where h is the Planck constant and p is the momentum of the charge carrier), and the carriers are confined in a finite potential well in the y -direction.

Figure 31b shows the energy levels in the quantum well derived in Appendix H. The values of E_n are shown as E_1, E_2, E_3 for electrons, $E_{hh1}, E_{hh2}, E_{hh3}$, for heavy holes,[§] and E_{lh1}, E_{lh2} for light holes.²¹ The usual parabolic forms for the conduction and valence band density of states have been replaced by a “staircase” representation of discrete levels (Fig. 31c). Since the density of states is constant rather than gradually increasing from zero, as in a conventional laser, there is a group of electrons of nearly the same energy available shown in Fig. 31d to recombine with a group of holes of nearly the same energy, for example, the level E_1 in the conduction band with the level E_{hh1} in the valence band. The sharper electron profile at the band edge, E_1 in this case, makes population inversion much easier to achieve, so that QW lasers offer significant improvement in laser performance, such as reduced threshold current, high output power, and high speed, compared with the conventional DH lasers. QW lasers made in GaAs/AlGaAs material systems have threshold current densities as low as 65 A/cm² and submilliampere threshold currents. These lasers operate at emission wavelengths around 0.9 μm .

At high current bias, more than one subbands are filled with injected carriers. The internal emission spectrum is thus much wider. The lasing wavelength, however, is also selected by other means such as the optical cavity length. So in a quantum-well laser, wavelength tuning can cover a wider range.

[§] In GaAs, the effective mass for heavy holes is $0.62 m_0$ and that for light holes is $0.07 m_0$.

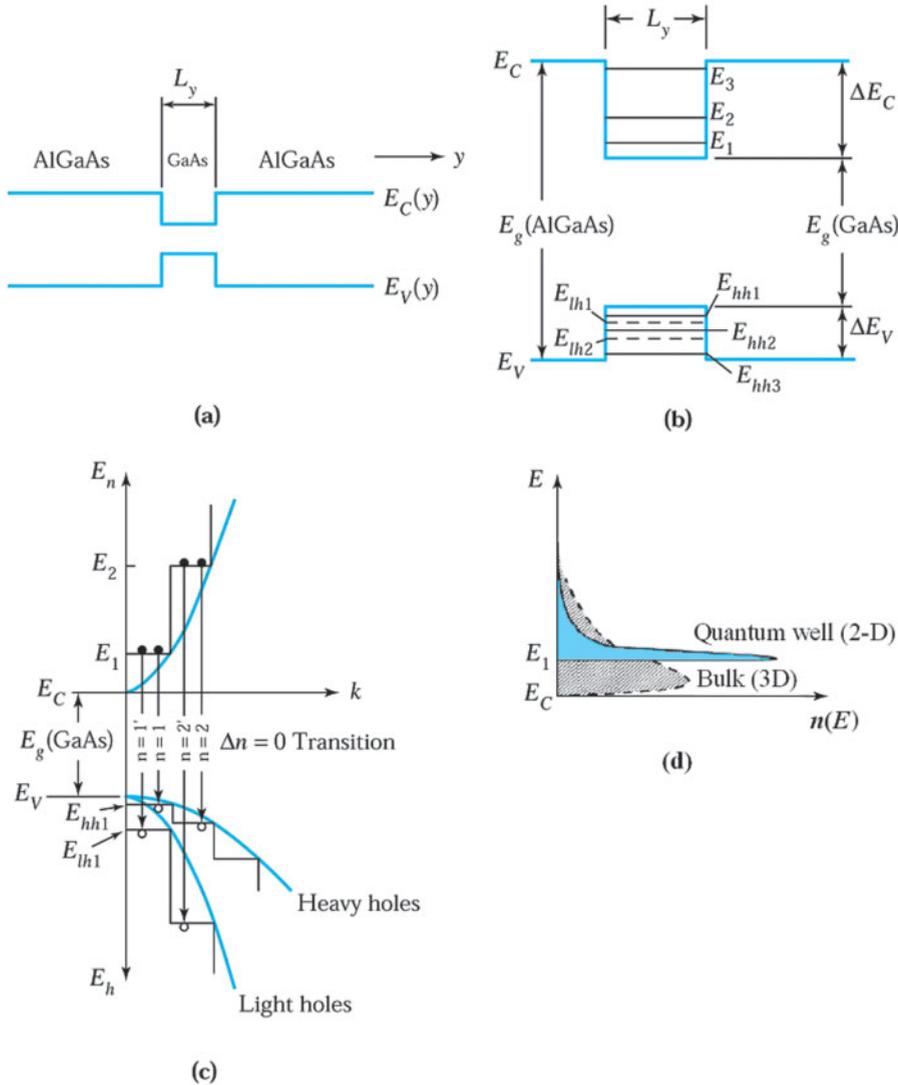


Fig. 31 The quantum-well (QW) laser: (a) single GaAs QW surrounded by AlGaAs, (b) discrete energy levels within the well, (c) density of states for electrons and holes within the well, and (d) electron concentration distribution.

9.4.6 Separate-Confinement Heterostructure MQW laser

One drawback of the thin active layer in a quantum-well laser is the poor optical confinement. This can be improved with multiple quantum wells stacked on the top of one another. Multiple-quantum-well lasers have higher quantum efficiency as well as higher output power. Single or multiple quantum wells can be incorporated in a separate-confinement heterostructure (SCH) scheme to improve optical confinement.

Figure 32a shows a schematic diagram of a separate-confinement-heterostructure (SCH) MQW laser for the 1.3 μm and 1.5 μm wavelength regions where four QWs of GaInAs with GaInAsP barrier layers are sandwiched between the InP cladding layers to form a waveguide with a step-index change.²³ These alloy compositions are chosen so that they are lattice matched to the InP substrate. The active region is composed of four 8 nm thick, undoped GaInAs QWs (with E_g of 0.75 eV) separated by 30 nm thick undoped GaInAsP barrier layers (with E_g of 0.95 eV).

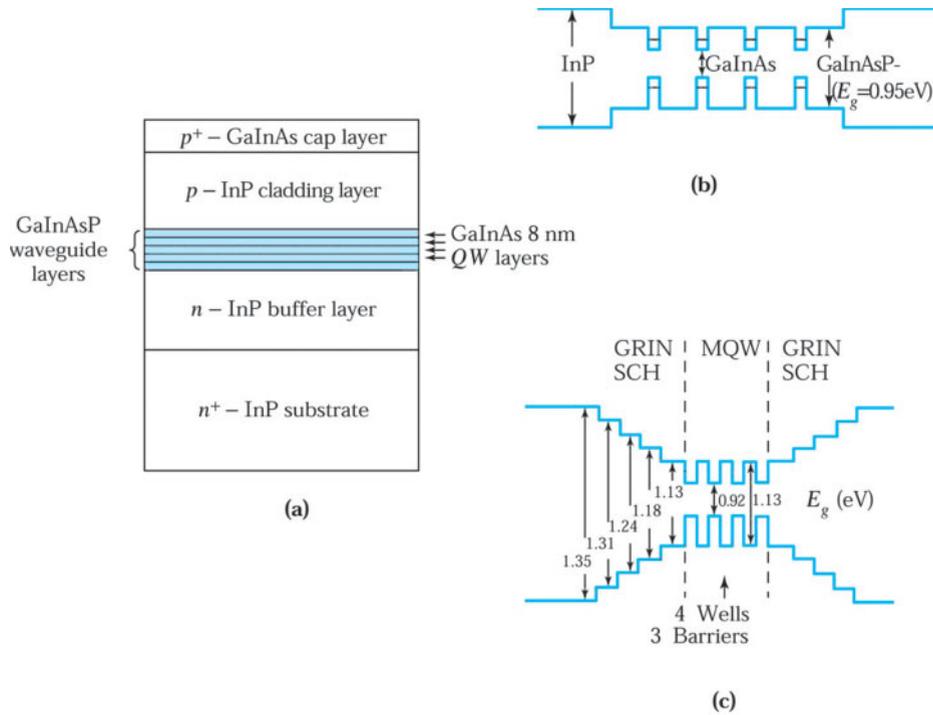


Fig. 32 (a) Schematic of the cross-section of an GaInAs/GaInAsP multiple-quantum-well laser structure. (b) Schematic of the bandgaps of the SCH-MQW layers shown in (a). (c) GRIN-SCH-MQW structure with thin layers of increasing bandgaps to approximate the graded-index change.²³

Figure 32b shows the corresponding band diagram of the active region. The n - and p -cladding InP layers are doped with sulfur (10^{18}cm^{-3}) and zinc (10^{17}cm^{-3}), respectively. A graded-index SCH (GRIN-SCH) is shown in Fig. 32c, in which a GRIN of the waveguide is accomplished by several small stepwise increases of the bandgap energies of multiple cladding layers. The GRIN-SCH structure confines both the carriers and the optical field more effectively than the SCH structure and, consequently, leads to an even lower threshold current density.

9.4.7 Quantum-Wire and Quantum-Dot lasers

In quantum-wire and quantum-dot lasers, the active regions are reduced to the de Broglie wavelength regime, into 1-D (one dimension) (wire) and 0-D (zero dimension) (island) formation. These wires and dots are placed between a p - n junction as shown in Fig. 33. To realize such small dimensions, the small active regions are mostly formed by epitaxial regrowth on specially processed surfaces (etched, cleaved, vicinal, or V-groove), or by a process called self-ordering after epitaxy. The advantages of these lasers are similar to the quantum-well laser. These advantages also stem from their respective densities of states. These densities of states give rise to the optical-gain spectra which are compared in Fig. 34.²⁴

The optical gains include those from regular 3-D (bulk) active layer down to quantum dots. As seen, the peak gains for quantum wires and quantum dots are progressively higher, and their shapes are sharper. These gain characteristics give low threshold current. The reduction of threshold current for different structures are summarized in Fig. 35,^{25,26} which also indicates their introduction in chronological order.

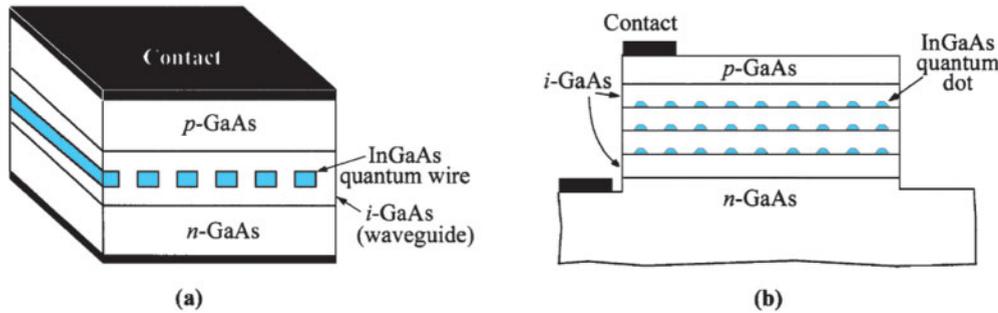


Fig. 33 Simplified schematic structures for (a) quantum-wire laser and (b) quantum-dot laser.

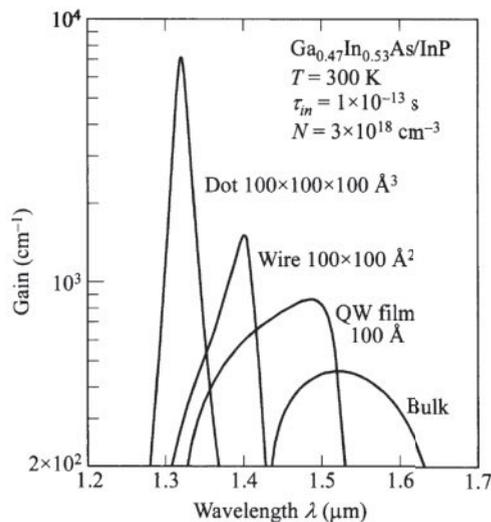


Fig. 34 Calculated optical gain vs. wavelength for different dimensionality. Note increase of peak gain and narrower spectrum as dimensionality is reduced.²⁴

9.4.8 Vertical-Cavity Surface-Emitting Laser (VCSEL)

So far the lasers discussed are edge-emitting so that the light output is parallel to the active layer. In a surface-emitting laser as shown in Fig. 36, however, the light output is orthogonal to the active layer and the semiconductor surface, whence the name vertical-cavity surface-emitting laser (VCSEL).²⁷ The VCSEL usually has an active layer formed by multiple quantum wells. The optical cavity is formed by two distributed-Bragg reflectors (DBRs) surrounding the active layer. These DBRs have a high reflectivity of more than 90%.

The high reflectivity is required since the optical gain per pass is small due to the small optical cavity compared to an edge-emitting laser. The benefits of a small optical cavity include low threshold current, and single-mode lasing, since the mode separation is wide (Eq. 29). Other advantages of the VCSEL are the realization of a 2-D laser array, ease of coupling light output to other media such as optical fiber and optical interconnect, compatibility with IC processing for integrated optics, high-volume and low-cost production, high speed, and on-wafer testing capability.

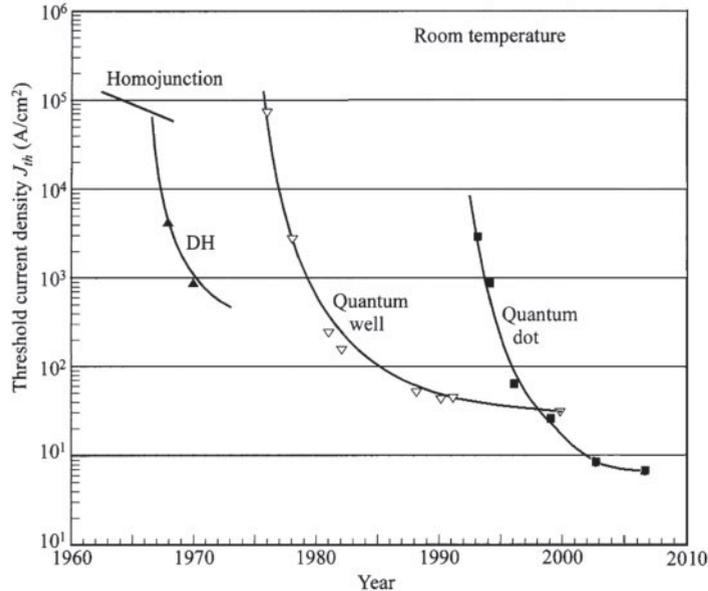


Fig. 35 Reduction of threshold current density from homojunction laser to DH, quantum-well, and quantum-dot lasers.^{25,26}

9.4.9 Quantum-Cascade Laser

The structure of the quantum cascade laser is shown in Fig. 37.²⁸ The active region is composed of multiple quantum wells (usually two or three quantum wells) or a *superlattice*, which create quantized subband energy levels in the conduction band. The electron transition between intersubband emits a photon with an energy much smaller than the energy gap. The quantum cascade laser is capable of lasing in long wavelengths, without the difficulties of very narrow-energy-gap semiconductors, which are much less stable and less developed. Wavelengths beyond $70\ \mu\text{m}$ have been achieved. Besides, the wavelength is tunable by varying the quantum-well thicknesses. The intersubband transition is the major difference from the interband transition in a conventional laser.

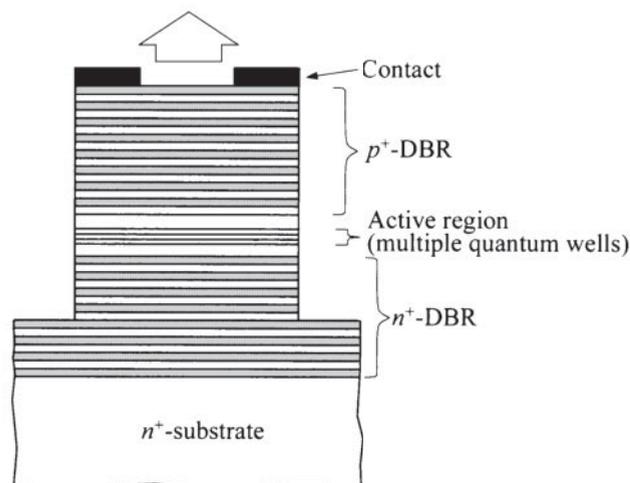


Fig. 36 Structure of a vertical-cavity surface-emitting laser (VCSEL).²⁷

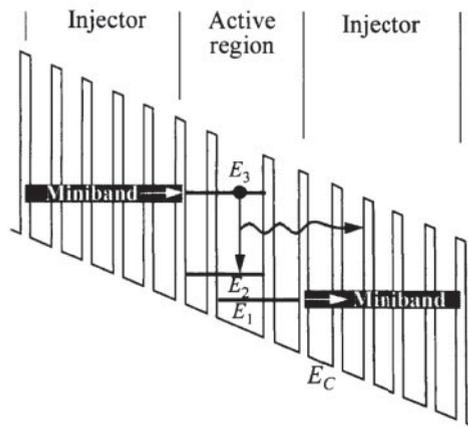


Fig. 37 Energy-band diagram showing conduction-band edge E_C of the quantum cascade laser under biasing condition. A period consists of an active region and a superlattice injector, and is repeated in series.²⁸

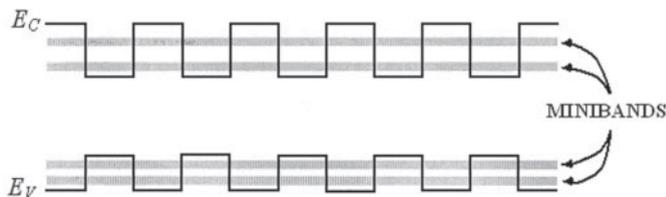


Fig. 38 Energy-band diagrams for heterostructure superlattice.

The difference between multiple quantum wells and a superlattice is that when quantum wells are separated from one another by thick barrier layers, there is no communication between them and this system can be described only as multiple quantum wells. However, when the barrier layers between them become thin enough that wavefunctions start to overlap, a heterostructure superlattice is formed (e.g., GaAs/Al_xGa_{1-x}As, with each layer 10 nm or less in thickness). The superlattice has two major differences from a multiple-quantum-well system: (1) the energy levels are continuous in space across the barrier, and (2) the discrete bands widen into minibands, as shown in Fig. 38. Since the continuous conduction band is divided into subbands, electrons no longer reside on the band edges E_C but on these subbands only.

The electron injector is composed of a *superlattice* with the miniband formed in the conduction band. Electrons in the injector are injected through resonant tunneling to the sublevel E_3 in the active region. (see Section 8.5 for resonant tunneling.) The radiative transition between E_3 and E_2 in the active region is responsible for lasing. Electrons in E_2 relax to E_1 and then tunnel to the miniband of the succeeding injector through resonant tunneling, or they can also tunnel directly from E_2 to the injector. Resonant tunneling is a very fast process so that the electron concentration in E_2 is always less than that in E_3 ; thus population inversion is maintained.

Design of the minibands plays a critical role and is dependent on the nonuniform thicknesses of the quantum wells. Notice that in Fig. 37 E_3 is not aligned to a miniband of the succeeding injector, so tunneling to the injector is blocked and high concentration at E_3 can be maintained. The design of the injector is also critical. Under bias, the miniband should remain flat for efficient resonant tunneling. This has to be done by careful tailoring of the injector superlattice with a special doping profile, thickness profile, or barrier profile.

The period, consisting of the active region plus the injector, is repeated many times (20-100), and this cascade scheme helps to establish a high external quantum efficiency and low threshold current since the same carrier can produce many photons. This phenomenon is not possible in a conventional laser. Due to the small transition energy, lower-temperature operation is required. Nevertheless, CW operation has been achieved at $\sim 150\text{K}$ and pulsed operation at room temperature.

► SUMMARY

The operation of photonic devices (LEDs and laser diodes) depends upon the emission of photons. Photons are emitted when charge carriers recombine.

LEDs are p - n junctions that can emit spontaneous radiation due to recombination of electrons and holes in a forward-biased junction. Visible LEDs can emit radiation with photon energies in the range of 1.8 to 2.8 eV, corresponding to wavelengths from 0.7 to 0.4 μm . They are used extensively for displays and various electronic instruments. By combining LEDs of different colors (i.e., red, green, and blue), we can form white LEDs that are useful for general illumination. Organic semiconductors can also be used for display applications. OLED is particularly useful for multicolor large-area flat-panel displays. Infrared LEDs can emit radiation with $h\nu < 1.8$ eV. They are used for opto-isolator and short-distance optical-fiber communication.

The laser diode is also a p - n junction operated under forward bias condition. However, the diode structure must provide the confinement of the carriers and the optical field so that the stimulated emission condition can be established. Laser diodes have evolved from the homojunction, to double heterojunction, to distributed feedback configuration, and to quantum-well structures. The main objectives are to lower the threshold current density and to have single-frequency operation. The laser diode is a key device in long-distance optical-fiber communication systems. It is also extensively used for video recording, high-speed printing, and optical reading.

► REFERENCES

1. H. C. Casey, Jr. and M. B. Panish, *Heterostructure Lasers*, Academic, New York, 1978.
2. H. Melchior, "Demodulation and Photodetection Techniques," in F. T. Arecchi and E. O. Schulz-Dubois, Eds., *Laser Handbook*, Vol. 1, North-Holland, Amsterdam, 1972.
3. R. H. Saul, T. P. Lee, and C. A. Burms, "Light-Emitting Diode Device Design," in R. K. Willardon and A. C. Bear, Eds., *Semiconductors and Semimetals*, Academic, New York, 1984.
4. S. Gage, D. Evans, M. Hodapp, and H. Sorensen, *Optoelectronic Application Manual*, McGraw-Hill, New York, 1977.
5. I. Schnitzer, E. Yablonovitch, C. Caneau, T. J. Gmitter, and A. Scherer, "30% external quantum efficiency from surface textured, thin-film light emitting diodes," *Appl. Phys. Lett.*, **63**, 2174 (1993).
6. M. G. Craford, "Recent Developments in LED Technology," *IEEE Trans. Electron Devices*, **ED-24**, 935 (1977).
7. W. O. Groves, A. H. Herzog, and M. G. Craford, "The Effect of Nitrogen Doping on GaAsP Electroluminescent Diodes," *Appl Phys. Lett.*, **19**, 184 (1971).
8. E. F. Schubert, *Light-Emitting Diodes*, 2nd edition, Cambridge, UK, 2006.
9. A. A. Bergh and P. J. Dean, *Light Emitting Diodes*, Clarendon, Oxford, 1976.
10. N. Bailey, "The Future of Organic Light-Emitting Diodes," *Inf. Disp.*, **16**, 12 (2000).
11. C. H. Chen, J. Shi, and E. W. Tang, "Recent Developments in Molecular Organic Electroluminescent Materials," *Macromol. Symp.*, **125**, 1 (1997).

12. L. S. Rohwer and A. M. Srivastava, "Development of Phosphors for LEDs," *Interface*, 36 (summer 2003).
13. S. E. Miller and A. G. Chynoweth, Eds., *Optical Fiber Communications*, Academic, New York, 1979.
14. T. Miya, Y. Terunuma, T. Hosaka, and T. Miyashita, "Ultimate Low-Loss Single Mode Fiber at 1.55 μm ," *Electron. Lett.*, **15**, 108 (1979).
15. W. T. Tsang, "High Speed Photonic Devices," in S. M. Sze, Ed., *High Speed Semiconductor Devices*, Wiley, New York, 1990.
16. O. Madelung, Ed., *Semiconductor-Group IV Elements and III-V Compounds*, Springer-Verlag, Berlin, 1991.
17. M. B. Panish, I. Hayashi, and S. Sumski, "Double-Heterostructure Injection Lasers with Room Temperature Threshold as Low as 2300 A/cm²," *Appl. Phys. Lett.*, **16**, 326 (1970).
18. T. E. Bell, "Single-Frequency Semiconductor Lasers," *IEEE Spectrum*, **20**, 38 (1983).
19. F. Stern, "Calculated Spectral Dependence of Gain in Excited GaAs," *J. Appl. Phys.*, **47**, 5328 (1976).
20. W. T. Tsang, R. A. Logan, and J. P. Van der Ziel, "Low-Current-Threshold Stripe-Buried-Heterostructure Laser with Self-Aligned Current Injection Stripes," *Appl. Phys. Lett.*, **34**, 644 (1979).
21. N. Holonyak, Jr., R. M. Kolbas, R. D. Dupuis, and P. D. Dapkus, "Quantum Well Heterostructure Laser," *IEEE J. Quant. Electron.*, **QE-16**, 170 (1980).
22. T. P. Lee, "High Speed Photonic Devices," in S. M. Sze, Ed., *Modern Semiconductor Device Physics*, Wiley Interscience, New York, 1998.
23. K. Kasukawa, Y. Imajo, and T. Makino, "1.3 μm GaInAsP/InP Buried Heterostructure Graded Index Separate Confinement Multiple Quantum Well Lasers Epitaxially Grown by MOCVD," *Electron. Lett.*, **25**, 104 (1989).
24. M. Asada Y. Miyamoto and Y. Suematsu, "Gain and the Threshold of Three-Dimensional Quantum-Box Laser," *IEEE J. Quantum Electron.*, **QE-22**, 1915 (1986).
25. N. N. Ledentsov, M. Grundmann, F. Heinrichsdorff, D. Bimberg, V. M. Ustinov, A. E. Zhukov, M. V. Maximov Z. I. Alferov, and J. A. Lott, "Quantum-Dot Heterostructure Lasers," *IEEE J. Selected Topics Quan. Elect.*, **6**, 439 (2000).
26. J. P. Reithmaier, "Quantum Dot Laser," Tutorial for WWW. BRIGHTER. EU, Lund (June 2007).
27. K. D. Choquett, "Vertical-Cavity Surface-Emitting Lasers: Light for the Information Age," *MRS Bulletin*, 507, (July 2002).
28. F. Capasso, R. Paiella, R. Martini, R. Colombelli, C. Gmachl, T. L. Myers, M. S. Taubman, R. M. Williams, C. G. Bethea, K. Unterrainer H. Y. Hwang, D. L. Sivco, A. Y. Cho, A. M. Sergent, H. C. Liu and E. A. Whittaker, "Quantum Cascade Lasers: Ultrahigh-Speed Operation, Optical Wireless Communication, Narrow Linewidth, and Far-Infrared Emission," *IEEE J. Quantum Electron.*, **QE-38**, 511 (2002).

► PROBLEMS (* DENOTES DIFFICULT PROBLEMS)

FOR SECTION 9.1 RADIATIVE TRANSITIONS AND OPTICAL ABSORPTION

- *1. A GaAs sample is illuminated with a light having a wavelength of $0.6\ \mu\text{m}$. The incident power is $15\ \text{mW}$. If one-third of the incident power is reflected and another third exits from the other end of the sample, what is the thickness of the sample? Find the thermal energy dissipated per second to the lattice.
2. The absorption coefficient is $4 \times 10^4\ \text{cm}^{-1}$ and surface reflectivity is 0.1 for a Si wafer illuminated with a monochromatic light having an $h\nu$ of $3\ \text{eV}$. Calculate the depth at which half the incident optical power has been absorbed in a material.

FOR SECTION 9.2 VARIOUS LIGHT-EMITTING DIODES

3. Calculate the spectral half-width at room temperature of an infrared LED of peak wavelength $550\ \text{nm}$.
4. The efficiency for electrical-to-optical conversion in a LED is given by $4 \bar{n}_1 \bar{n}_2 (1 - \cos\theta_c) / (\bar{n}_1 + \bar{n}_2)^2$, where n_1 and n_2 are the refractive index of air and the semiconductor, respectively, and θ_c is the critical angle. Find the efficiency of an $\text{Al}_{0.3}\text{Ga}_{0.7}\text{As}$ LED operated at $0.898\ \mu\text{m}$.
5. Assume that the radiative lifetime τ_r is given by $\tau_r = 10^9/N$ seconds, where N is the semiconductor doping in cm^{-3} and the nonradiative τ_{nr} is equal to $10^{-7}\ \text{s}$. Find the cutoff frequency of an LED having a doping of $10^{19}\ \text{cm}^{-3}$.
6. Calculate the 3 dB (half-power) frequency of the LED in Prob. 5.
7. Calculate the Fresnel transmission coefficient for an $\text{Al}_{0.3}\text{Ga}_{0.7}\text{As}$ LED of refractive index 3.38 operated at $0.898\ \mu\text{m}$ as in Prob. 4.
8. What is the optical output power of an $\text{Al}_{0.3}\text{Ga}_{0.7}\text{As}$ LED operated at a forward voltage of $1.8\ \text{V}$ and $80\ \text{mA}$ if the internal quantum efficiency is 0.716 and the extraction efficiency is 1 ? Assume that this power is radiated symmetrically about the junction, all light toward the top region eventually leaves the surface by internal reflection, and all light toward the bottom region is lost. Calculate the amount of light reaching the surface of the p -region if it is $3\ \mu\text{m}$ thick and the adsorption coefficient is $5 \times 10^3\ \text{cm}^{-1}$.
9. What is the optical output power of the LED in Prob. 8 deposited a dielectric of refractive index of 1.6 on its surface?
10. The bandgap of GaAs is $1.42\ \text{eV}$ at $300\ \text{K}$ and decreases with temperature as shown in Fig. 28 in Chapter 1. From measurement, the change of emitted wavelength of GaAs LED of $2.8\ \text{nm}$ corresponds to the temperature change of $10\ ^\circ\text{C}$. Derive dE_g/dT .
11. For a GaAs LED operated at $0.8\ \mu\text{m}$, calculate (a) the fraction of photons to be emitted from the junction into air if we take total internal reflection into account, and (b) if we also consider Fresnel loss.

FOR SECTION 9.4 SEMICONDUCTOR LASERS

12. An InGaAsP Fabry-Perot laser operating at a wavelength of $1.33\ \mu\text{m}$ has a cavity length of $300\ \mu\text{m}$. The index of refraction of InGaAsP is 3.39 . (a) What is the mirror loss expressed in cm^{-1} ? (b) If one of the laser facets is coated to produce 90% reflectivity, how much threshold current reduction (as a percentage) can be expected, assuming $\alpha = 10\ \text{cm}^{-1}$?
13. (a) For an InGaAsP laser operating at a wavelength of $1.3\ \mu\text{m}$, calculate the mode spacing in nanometer for a cavity of $300\ \mu\text{m}$, assuming the group refractive index is 3.4 . (b) Express the mode spacing obtained above in GHz.

14. From Fig. 26 the threshold current densities for conventional p - n junction and DH lasers with Fabry-Perot cavity of $300\ \mu\text{m}$ are 5×10^4 and $10^3\ \text{A}/\text{cm}^2$. Calculate the threshold current density for a conventional p - n junction laser and a stripe-geometry DH laser with a stripe width $20\ \mu\text{m}$.
15. Calculate the confinement factor for a GaAs laser with active region thickness $1\ \mu\text{m}$, refractive index 3.6, and critical angle at the active-to-nonactive boundary of 84° . Assume the C constant to be $8 \times 10^7\ \text{m}^{-1}$. Repeat the calculation for a GaAs/AlGaAs DH laser, where all the factors remain unchanged except for the critical angle, which is now 78° .
16. Derive Eq. 29 for the separation $\Delta\lambda$ between the allowed modes in the longitudinal direction. For a GaAs laser diode operated at $\lambda = 0.89\ \mu\text{m}$, with $\bar{n}_1 = 3.58$, $L = 300\ \mu\text{m}$, and $d\bar{n}_1/d\lambda = 2.5\ \mu\text{m}^{-1}$, find $\Delta\lambda$.
17. Calculate the gain coefficient for the two cases in Prob. 15 if the cavity length is $100\ \mu\text{m}$, the absorption coefficient is $10^4\ \text{m}^{-1}$, and the end mirrors are cleaved. How much shorter can the cavity be and still produce the same gain if one end-mirror is metallized to produce a reflectivity of 0.99?
18. Calculate the threshold current for the two cases in Prob. 15 if one end-mirror's reflectivity is 0.99. The cavity width is $5\ \mu\text{m}$ and the gain factor for case 1 is $0.1\ \text{cm}^3\text{A}^{-1}$.
- *19. For a DFB laser with a cavity length of $300\ \mu\text{m}$, a material refractive index of 3.4, and an oscillating wavelength of $1.33\ \mu\text{m}$, find the Bragg wavelength and grating periodicity.

The oscillating wavelength λ_o is given by $\lambda_o = \lambda_B \pm \frac{(m + \frac{1}{2})\lambda_B^2}{2\bar{n}L}$, where m is an integer.

20. For high-temperature laser operation, it is important to have a low-temperature coefficient of the threshold current $\xi = (dI_{th}/dT)/I_{th}$. What is the coefficient ξ for the laser shown in Fig. 29? If $T_o = 50^\circ\text{C}$, is this laser better or worse for high-temperature operation?

Photodetectors and Solar Cells

- ▶ 10.1 PHOTODETECTORS
 - ▶ 10.2 SOLAR CELLS
 - ▶ 10.3 SILICON AND COMPOUND-SEMICONDUCTOR SOLAR CELLS
 - ▶ 10.4 THIRD-GENERATION SOLAR CELLS
 - ▶ 10.5 OPTICAL CONCENTRATION
 - ▶ SUMMARY
-

Photodetectors are semiconductor devices that electrically detect optical signals. At its operating wavelength, the photodetector should have high sensitivity, high response speed, minimum noise, small size, low voltage, and high reliability under operating conditions. *Solar cells* are used to generate power from the sunlight and have some similarities with photodetectors. The main differences between them are device area, operating frequency, and optical source.

Specifically, we cover the following topics:

- Absorption of photons to create electron-hole pairs for photodetectors.
- Structures of some important photodetectors.
- Absorption of photons to convert them to electrical energy from solar cells.
- Structures of some important solar cells.

▶ 10.1 PHOTODETECTORS

Photodetectors are semiconductor devices that can convert optical signals into electrical signals. The operation of a photodetector involves three steps: carrier generation by incident light, carrier transport and/or multiplication by whatever current-gain mechanism may be available, and interaction of the current with the external circuit to provide the output signal.

Photodetectors have a broad range of applications, including infrared sensors in optoisolators and detectors for optical-fiber communications. For these applications, the photodetectors must have high sensitivity at the operating wavelengths, high response speed, and low noise. In addition, the photodetector should be compact, use low biasing voltages or currents, and be reliable under the required operating conditions.

10.1.1 Photoconductor

A photoconductor consists simply of a slab of semiconductor with ohmic contacts at each end of the slab as shown in Fig. 1*a*, and a corresponding layout that consists of interdigitated contacts shown in Fig. 1*b*. When incident light falls on the surface of the photoconductor, electron-hole pairs are generated either by band-to-band transition (intrinsic) or by transitions involving forbidden-gap energy levels (extrinsic), resulting in an increase in conductivity.

For the intrinsic photoconductor, the conductivity is given by

$$\sigma = q(\mu_n n + \mu_p p), \tag{1}$$

and the increase in conductivity under illumination is due mainly to the increase in the number of carriers. The long-wavelength cutoff for an intrinsic photoconductor is given by Eq. 9, Chapter 9. For the extrinsic photoconductor, photoexcitation may occur between the band edge and an energy level in the energy gap. In this case, the long-wavelength cutoff is determined by the depth of the forbidden-gap energy level.

Consider the operation of a photoconductor under illumination. At time zero, the number of carriers generated in a unit volume by a given photon flux is n_0 . At a later time t , the number of carriers $n(t)$ in the same volume decays by recombination as

$$n = n_0 \exp\left(-\frac{t}{\tau}\right), \tag{2}$$

where τ is the carrier lifetime. From Eq. 2 the recombination rate is

$$\left|\frac{dn}{dt}\right| = \frac{1}{\tau} n_0 \exp\left(-\frac{t}{\tau}\right) = \frac{n}{\tau}. \tag{3}$$

If we assume a steady flow of photon flux impinging uniformly on the surface of a photoconductor (Fig. 1a) with an area $A = WL$, the total number of photons arriving at the surface is $(P_{opt}/h\nu)$ per unit time, where P_{opt} is the incident optical power and $h\nu$ is the photon energy. At steady state, the carrier-generation rate G must be equal to the recombination rate n/τ . If the detector thickness D is much larger than the light penetration depth $1/\alpha$, the total steady-state carrier-generation rate per unit volume is

$$G = \frac{n}{\tau} = \frac{\eta(P_{opt}/h\nu)}{WLD}, \tag{4}$$

where η is the quantum efficiency, the number of carriers generated per photon, and n is the carrier density, the number of carriers per unit volume. The photocurrent flowing between the electrodes is

$$I_p = (\sigma \mathcal{E})WD = (q\mu_n n \mathcal{E})WD = (qn v_d)WD, \tag{5}$$

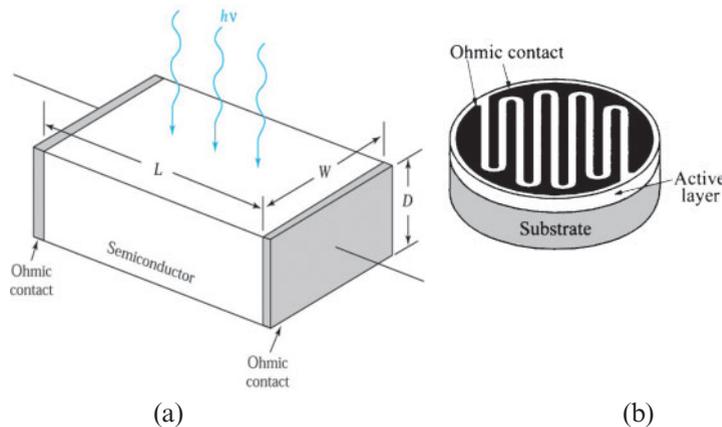


Fig. 1 (a) Schematic diagram of a photoconductor that consists of a slab of semiconductor and a contact at each end. (b) Typical layout consists of interdigitated contacts with a small gap.

where \mathcal{E} is the electric field inside the photocoductor and v_d is the carrier drift velocity. Substituting n in Eq. 4 into Eq. 5 gives

$$I_p = q \left(\eta \frac{P_{opt}}{h\nu} \right) \cdot \left(\frac{\mu_n \tau \mathcal{E}}{L} \right). \quad (6)$$

If we define the primary photocurrent as

$$I_{ph} = q \left(\eta \frac{P_{opt}}{h\nu} \right), \quad (7)$$

the photocurrent gain from Eq. 6 is

$$\text{Gain} \equiv \frac{I_p}{I_{ph}} = \frac{\mu_n \tau \mathcal{E}}{L} \frac{\tau}{t_r}, \quad (8)$$

where $t_r \equiv L/v_d = L/\mu_n \mathcal{E}$ is the carrier transit time. The gain depends on the ratio of carrier lifetime to the transit time.

► EXAMPLE 1

Calculate the photocurrent and gain when 5×10^{12} photons/s are arriving at the surface of a photoconductor of $\eta = 0.8$. The minority carrier lifetime is 0.5 ns, and the device has $\mu_n = 2500 \text{ cm}^2/\text{V}\cdot\text{s}$, $\mathcal{E} = 5000 \text{ V/cm}$, and $L = 10 \text{ }\mu\text{m}$.

SOLUTION From Eq. 6,

$$\begin{aligned} I_p &= q \left(0.8 \times 5 \times 10^{12} \text{ photons/s} \right) \cdot \left(\frac{2500 \text{ cm}^2/\text{V}\cdot\text{s} \cdot 5 \times 10^{-10} \text{ s} \cdot 5000 \text{ V/cm}}{10 \times 10^{-4} \text{ cm}} \right) \\ &= 4 \times 10^{-6} \text{ A} = 4 \text{ }\mu\text{A} \end{aligned}$$

and from Eq. 8,

$$\text{Gain} = \frac{\mu_n \tau \mathcal{E}}{L} = \frac{2500 \cdot 5 \times 10^{-10} \cdot 5000}{10 \times 10^{-4}} = 6.25. \quad \blacktriangleleft$$

For a sample with long minority-carrier lifetime and short electrode spacing, the gain can be substantially greater than unity. Gains as high as 10^6 can be obtained from some photoconductors. The response time of a photoconductor is determined by the transit time t_r . To achieve short transit time requires that we use small electrode spacing and a high electric field. The response times of photoconductors cover a wide range, from 10^{-3} to 10^{-10} seconds. They are extensively used for infrared detection, especially for wavelengths greater than a few micrometers.

10.1.2 Photodiode

A photodiode is basically a p - n junction operated under reverse bias. The space-charge and the electric-field distributions are similar to those in Fig. 6 in Chapter 3 except under reverse bias. Note that the electric-field distribution is nonuniform and the maximum field is at the junction. When an optical signal penetrates into the depletion region of the photodiode, the electric field in the depletion region serves to separate the photogenerated EHPs (electron-hole pairs) and an electric current, called photocurrent I_p , flows in the external circuit. The photogenerated holes drift in the depletion region, diffuse into the neutral p region, and then combine with

electrons entered from the negative electrode. Similarly, photogenerated electrons drift in the opposite direction. When an optical signal penetrates within a diffusion length outside the depletion region, the photogenerated carriers will diffuse into the depletion region and drift across the depletion region to the other side. These neutral regions can be regarded as resistive extensions of electrodes to the depletion region. The photocurrent depends on the number of photogenerated EHPs and the drift velocities of the carriers. It should be noted that the photocurrent in the external circuit is due only to the flow of electrons, even though there is electron and hole drift in the depletion region.

For high-frequency operation, the depletion region must be kept thin to reduce the transit time. On the other hand, to increase the quantum efficiency the depletion layer must be sufficiently thick to allow a large fraction of the incident light to be absorbed. Thus, there is a trade-off between the response speed and quantum efficiency.

Quantum Efficiency

The quantum efficiency, as mentioned above, is the number of EHPs generated for each incident photon:

$$\eta = \left(\frac{I_p}{q} \right) \cdot \left(\frac{P_{opt}}{h\nu} \right)^{-1}, \quad (9)$$

where I_p is the photogenerated current from the absorption of incident optical power P_{opt} at a wavelength λ (corresponding to a photon energy $h\nu$) and is known more specifically as the external quantum efficiency. The internal quantum efficiency is defined as the photogenerated number of EHPs per absorbed photon. One of the key factors that determine η is the absorption coefficient α (Fig. 5, Chapter 9). Since α is a strong function of the wavelength, the wavelength range in which appreciable photocurrent can be generated is limited. The long-wavelength cutoff λ_c is established by the bandgap (Eq. 9, Chapter 9) and is about 1.8 μm for germanium and 1.1 μm for silicon. For wavelengths longer than λ_c , the values of α are too small to give appreciable band-to-band absorption. For wavelengths much shorter than λ_c , the values of α are too large ($\sim 10^5 \text{ cm}^{-1}$), and hence the radiation is mostly absorbed very near the surface where recombination time is short. Therefore, the photocarriers can recombine before they can be collected in the depletion region of p - n junction.

The photogenerated carriers in the depletion region may disappear by recombination or by trapping without contributing to the photocurrent. The quantum efficiency is always less than unity. It depends on the absorption coefficient and the device structure. The quantum efficiency can be increased by reducing the surface reflection on the device to increase the absorption in the depletion region, and by preventing the recombination or trapping of carriers through improving material and device quality.

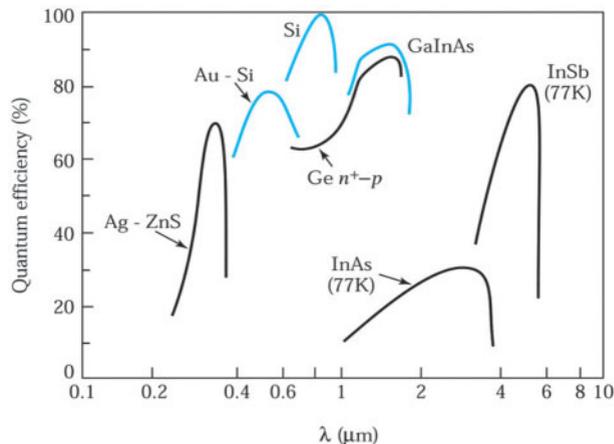


Fig. 2 Quantum efficiency versus wavelength for various photodetectors.^{1,2}

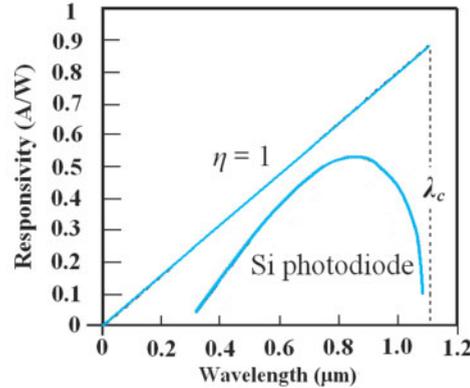


Fig. 3 Responsivity vs. wavelength for an ideal photodiode with $\eta = 1$ and for a typical commercial Si photodiode.

Figure 2 shows typical plots of quantum efficiency versus wavelength for some high-speed photodiodes.^{1,2} Note that in the ultraviolet and visible region, metal-semiconductor photodiodes (discussed in Section 10.1.4) show good quantum efficiencies. In the near-infrared region, silicon photodiodes (with an antireflection coating) can reach 100% quantum efficiency near the 0.8- to 0.9- μm region. In the 1.0- to 1.6- μm region, germanium photodiodes and Group III-V photodiodes (e.g., GaInAs) have shown high quantum efficiencies. For even longer wavelengths, photodiodes are cooled (e.g., to 77 K) for high-efficiency operation.

Responsivity

The responsivity \mathcal{R} of a photodiode is defined as the generated photocurrent (I_p) per incident optical power (P_{opt}). \mathcal{R} is also called the *spectral responsivity* or *radiant sensitivity*:

$$\mathcal{R} = I_p / P_{opt} \quad (10)$$

From the definition of quantum efficiency, we have

$$\mathcal{R} = I_p / P_{opt} = \eta q / h\nu = \eta q \lambda / hc \quad (11)$$

If a photodiode has an ideal quantum efficiency of 100%, then \mathcal{R} should be linearly proportional to the wavelength. In practice, the relationship of \mathcal{R} and λ is shown in Fig. 3. The quantum efficiency limits the responsivity below the ideal photodiode.

Response Speed

The response speed is limited by three factors: (1) diffusion of carriers, (2) drift time in the depletion region, and (3) capacitance of the depletion region. Carriers generated outside the depletion region must diffuse to the junction, resulting in considerable time delay. To minimize the diffusion effect, the junction should be formed very close to the surface. The greatest amount of light will be absorbed when the depletion region is wide. However, the depletion layer must not be too wide or transit time effects will limit the frequency response. It also should not be too thin, or excessive capacitance C will result in a large RC time constant, where R is the load resistance. The optimal compromise is the width at which the depletion layer transit time is approximately one half the modulation period. For example, for a modulation frequency of 2 GHz, the optimal depletion-layer thickness in silicon (with a saturation velocity of 10^7 cm/s) is about 25 μm .

10.1.3 *p-i-n* Photodiode

As described above, the *p-n* junction photodiode has two major drawbacks. First, the junction capacitance is not sufficiently small, due to the small depletion layer width. For example, the depletion layer width is below 1 μm for a p^+n silicon junction as in Ex. 2 in Chapter 3. It contributes a large RC time constant so that the photodiode cannot operate at high modulation frequencies. In addition, its depletion layer is not sufficiently wide to make the penetration depth greater than the depletion layer width at long wavelengths. The penetration depth is about 33 μm , for example, at the wavelength 900 nm shown in Fig. 5, Chapter 9. Most incident photons are absorbed outside the depletion region where there is no field to separate the EHPs.

The *p-i-n* (*p-intrinsic-n*) photodiode is one of the most common photodetectors, because the depletion layer thickness (the intrinsic layer) can be tailored to optimize the quantum efficiency and frequency response. The *i*-layer thickness is typically 5~50 μm depending on the particular application. The intrinsic *i*-layer in a *p-i-n* photodiode is completely depleted. The junction capacitance is sufficiently small due to the large depletion layer width to make the *p-i-n* photodiode operate at high modulation frequencies. Its depletion layer is also wide enough to have a large absorption in the depletion layer at long wavelengths.

Figure 4a shows a cross section of a *p-i-n* photodiode that has an antireflection coating to increase quantum efficiency. The surface reflection of the incident light from air ($n = 1$) into semiconductor silicon ($n = 3.5$) is about 0.31, from Eq. 22 in Chapter 9. This means that 31% of incident light is reflected and is not available for conversion to electrical energy. Covering the surface with an antireflection coating with a refractive index $n = (n_{\text{Si}})^{1/2}$ minimizes the total reflection. Si_3N_4 with $n = 1.9$ is a good choice. Figure 4b shows the energy band diagram of the *p-i-n* diode under reverse bias condition. The conduction band decreases linearly with distance and the electric field is uniform in the *i*-layer.

The optical absorption is shown in Fig. 4c. The *p-i-n* structure is designed so that almost complete optical absorption occurs over the *i*-layer. The EHPs produced in the depletion region or within a diffusion length of it from light absorption will eventually be separated by the electric field and a current flows in the external circuit as carriers drift across the depletion layer.

Generally, the response time is limited by the drift time of the slowest photogenerated carriers, holes, across the width of *i*-layer. A narrower *i*-layer improves the response time but decreases the quantity of absorbed photons and hence reduces the responsivity. To increase the response speed, i.e., to reduce the drift time, we have to increase the reverse bias. There is therefore a trade-off between speed and responsivity.

In practice, the *i*-layer will have a slight background doping. The structure is more like p^+-i-n^+ or p^+-v-n^+ mentioned in Fig. 27 in Chapter 3. The field is not uniform across the *i*-layer. As an approximation, we can still consider it as a *p-i-n* structure.

► EXAMPLE 2

On reaching the surface of the semiconductor, the incident optical power P_0 will have its level reduced to $P_0(1 - R)$ on entering the material, where R is the reflection coefficient. On passing through the semiconductor the light will be absorbed, and so at any depth x the amount of residual optical power $P(x)$ is given by $P(x) = P_0(1 - R) \exp(-\alpha x)$. For $\alpha = 10^4 \text{ cm}^{-1}$ and $R = 0.1$, calculate the depth at which half the incident optical power has been absorbed in a material.

SOLUTION

$$x = \frac{-1}{\alpha} \ln \left[\frac{P(x)}{P_0(1 - R)} \right] = -10^{-4} \cdot \ln \left(\frac{1}{2 \times 0.9} \right) \text{ cm}$$

$$= 0.59 \mu\text{m} \quad \blacktriangleleft$$

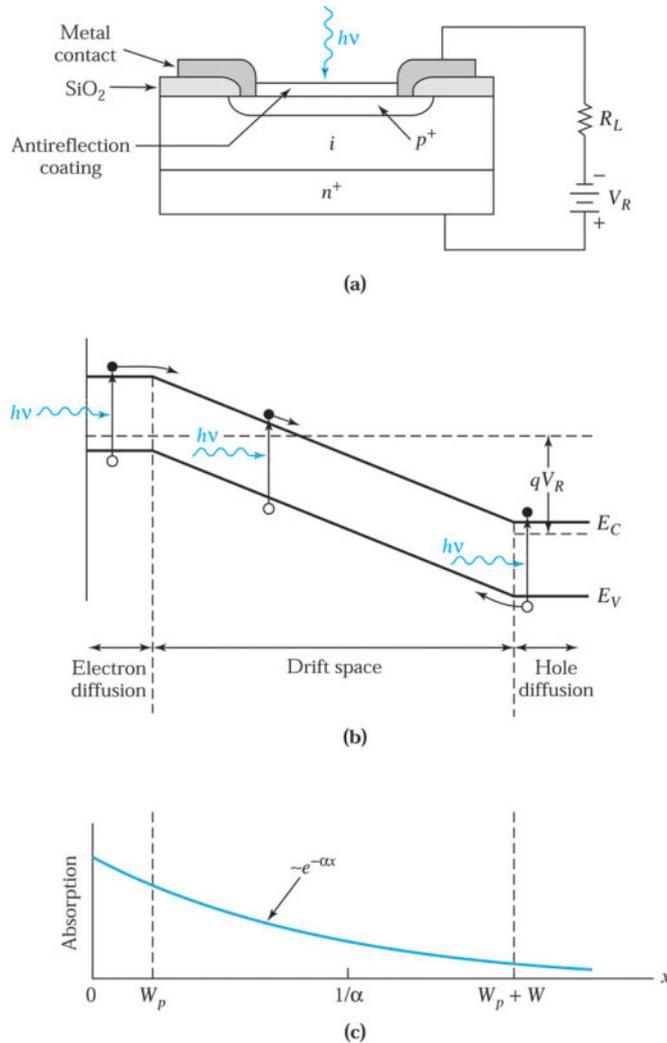


Fig. 4 Operation of a $p-i-n$ photodiode. (a) Cross-sectional view of a $p-i-n$ photodiode. (b) Energy band diagram under reverse bias. (c) Carrier absorption characteristics.

► EXAMPLE 3

The diameter of the optical receiving area of a Si $p-i-n$ photodiode is 0.06 cm. It is illuminated with an incident optical intensity of 0.2 mW/cm^2 at wavelength 800 nm to generate a photocurrent of $3 \times 10^{-4} \text{ mA}$. What are the responsivity and quantum efficiency of the $p-i-n$ photodiode at 800 nm?

SOLUTION

The incident optical intensity is 0.2 mW/cm^2 and the diameter of optical receiving area is 0.06 cm. Thus, the incident power is

$$P_{opt} = \pi (0.03 \text{ cm})^2 \times 0.2 \text{ mW/cm}^2 = 5.6 \times 10^{-4} \text{ mW}$$

The responsivity is

$$\mathcal{R} = I_p / P_{opt} = 3 \times 10^{-4} \text{ mA} / 5.6 \times 10^{-4} \text{ mW} = 0.54 \text{ A/W}$$

The quantum efficiency is

$$\eta = \mathcal{R}(hc/q\lambda) = 0.54 \text{ A/W} (6.62 \times 10^{-34} \text{ J-s})(3 \times 10^8 \text{ m/s}) / (1.6 \times 10^{-19} \text{ C})(80 \times 10^{-9} \text{ m}) = 0.84 = 84\% .$$

10.1.4 Metal-Semiconductor Photodiode

The construction of a high-speed metal-semiconductor (M-S) photodiode is shown in Fig. 5. To avoid large reflection and absorption losses when the diode is illuminated through the metal contact, the metal film must be very thin ($\sim 10 \text{ nm}$) and an antireflection coating must be used. Metal-semiconductor (M-S) photodiodes are particularly useful in the ultraviolet- and visible-light regions. In these regions the absorption coefficients, α , in most common semiconductors are very high, of the order of 10^5 cm^{-1} or more, which corresponds to an effective absorption length $1/\alpha$ of $0.1 \mu\text{m}$ or less. It is possible to choose a metal and an antireflection coating so that a large fraction of the incident radiation will be absorbed near the surface of the semiconductor. As an example, for a gold-silicon photodetector having 10 nm gold and 50 nm zinc sulfide as the antireflection coating, more than 95% of the incident light with $\lambda = 0.6328 \mu\text{m}$ (helium-neon laser wavelength, red light) will be transmitted into the silicon substrate.

The M-S photodiode can be operated in two modes, depending on the photon energy. For $h\nu > E_g$ (Fig. 6a), the radiation produces EHPs in the semiconductor, and the general characteristics of the M-S photodiode are very similar to those of a $p-i-n$ photodiode. For smaller photon energy (longer wavelength) $q\phi_B < h\nu < E_g$ (Fig. 6b), the photoexcited electrons in the metal can surmount the barrier and be collected by the semiconductor. This process is called *internal photoemission* and has been used extensively to determine the Schottky-barrier height and to study the hot electron transport in metal films.

When a Schottky-barrier diode is scanned with light of variable wavelength, Fig. 6c shows that the quantum efficiency has a threshold of $q\phi_B$ that increases with the photon energy. When the photon energy reaches the energy-gap value, the quantum efficiency jumps to a much higher value. In practical applications, however, the internal photoemission has typical quantum efficiencies of only less than 1%.

For detectors with internal photoemission, it is more efficient to direct the incoming light through the substrate. Since the barrier height is always smaller than the energy gap, light with $q\phi_B < h\nu < E_g$ is not absorbed in the semiconductor, and the intensity is not reduced at the metal/semiconductor interface. The metal layer, in this case, can be thicker for easier thickness control and to minimize series resistance. For Si devices, options are available using silicides in place of the metal. A silicide usually has a more reproducible interface since it is formed by reacting metal with Si so that the new interface is never exposed. Common silicides used for this purpose are PtSi, Pd₂Si, and IrSi. Another advantage of a Schottky-barrier diode is that it does not require high-temperature processing for diffusion or implantation annealing.

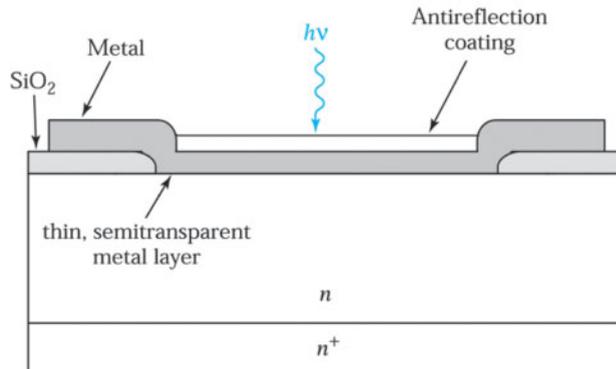


Fig. 5 Metal-semiconductor photodiode.

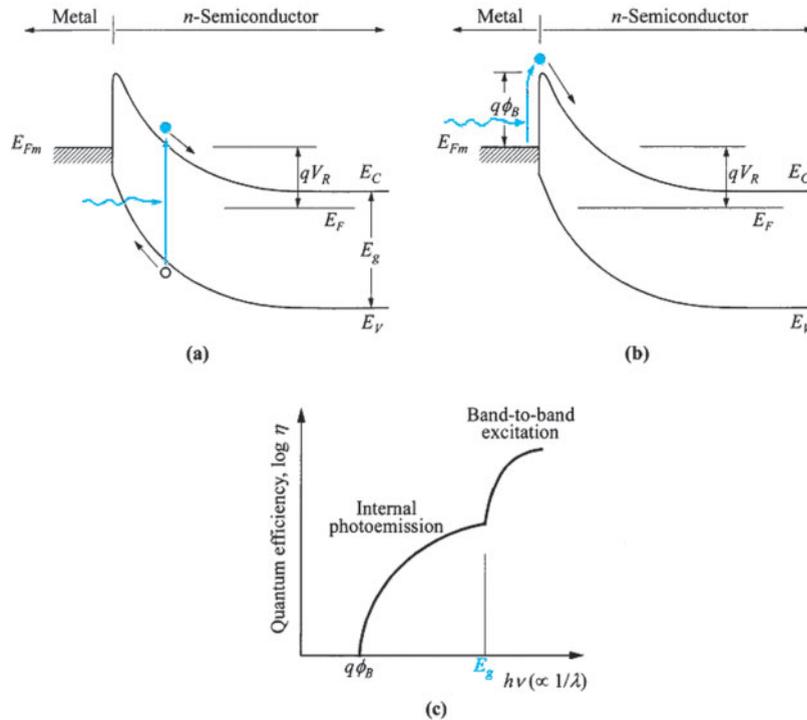


Fig. 6 (a) Band-to-band excitation of electron-hole pair ($h\nu > E_g$). (b) Internal photoemission of excited electrons from metal to semiconductor ($q\phi_B < h\nu < E_g$). (c) Quantum efficiency as a function of wavelength showing both processes.

10.1.5 Avalanche Photodiode

An avalanche photodiode (APD) is operated under a reverse-bias voltage that is sufficient to enable avalanche multiplication. The multiplication results in internal current gain and the device can respond to light modulated at frequencies as high as microwave frequencies.

One important consideration in the design of an APD is the need to minimize avalanche noise. The avalanche noise arises from the random nature of the avalanche multiplication process, in which every electron-hole pair generated at a given distance in the depletion region does not experience the same multiplication. The avalanche noise depends on the ratio of the ionization coefficients α_p/α_n ; the smaller the ratio, the smaller the avalanche noise. This is because when $\alpha_p = \alpha_n$, each incident photocarrier results in three carriers in the multiplying region: the primary carrier and its secondary hole and electron. A fluctuation that changes the number of carriers by one represents a large percentage change, and the noise will be large. On the other hand, if one of the ionization coefficients approaches zero (e.g., $\alpha_p \rightarrow 0$), each incident photocarrier can result in a large number of carriers in the multiplication region. In this case, a fluctuation of one carrier is a relatively insignificant perturbation. To minimize the avalanche noise, we should use semiconductors with a large difference in α_p and α_n . The noise factor is given by

$$F = M \left(\frac{\alpha_p}{\alpha_n} \right) + \left(2 - \frac{1}{M} \right) \left(1 - \frac{\alpha_p}{\alpha_n} \right) \quad (12)$$

where M is the multiplication factor. We can see from Eq. 12 that when $\alpha_p = \alpha_n$, the noise factor has a maximum value of M ; while for $\alpha_p/\alpha_n = 0$ and for a large M , the minimum noise factor is 2.

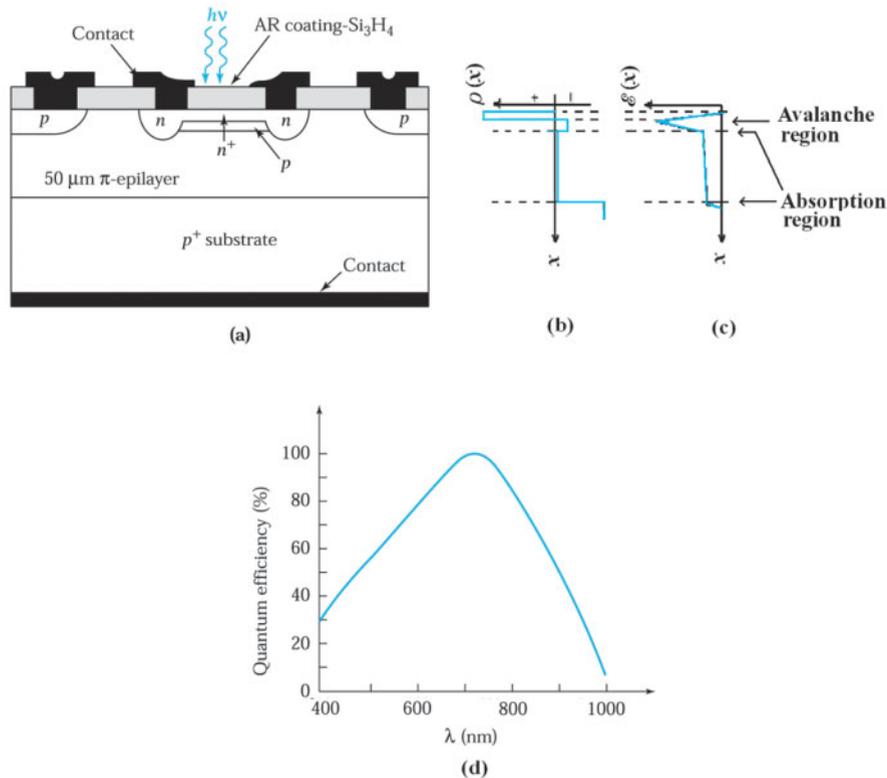


Fig. 7 A typical silicon avalanche photodiode. (a) Device structure. (b) Space charge distribution. (c) Electric-field distribution. (d) Quantum efficiency.

Figure 7a shows the structure of typical silicon APD having a n^+ - p - π - p^+ doping profile (π is a lightly doped p -region). The n^+ -side is thin and is illuminated through a window. There are three p -type layers p - π - p^+ with different doping concentrations next to the n^+ -layer. The net space charge distribution is shown in Fig. 7b. The field distribution across the diode is shown in Fig. 7c. The maximum electric field is at the n^+ - p junction, and then decreases slowly through the p -layer. It decreases slightly through the π -layer due to the small net space charge density. The field vanishes at the end of the narrow depletion layer in the p^+ -side. The diode is reverse biased to increase the fields in the depletion region. Under zero bias the depletion width in the p -region does not normally extend to the π -layer. Under sufficient reverse bias, the depletion width in the p -region can be widened to reach through the π -layer. The absorption of photons and hence EHPs generation is mainly in the π -layer due to the very thin n^+ and p layers. The electrons and holes drift in the π -layer at saturation velocities. When the electrons reach the p -layer, they experience high fields and acquire sufficient kinetic energy to cause avalanche and a large number of EHPs can be generated. This internal gain can result in a quantum efficiency in excess of unity.

The photogeneration is in the π -layer and the avalanche is in the p -layer. The advantage of the separation of photogeneration region and avalanche region is that the photogenerated electrons drift into the avalanche region but not the photogenerated holes in Fig. 7. The avalanche caused by electrons with the higher impact ionization efficiency has minimum noise.

There is an n -type doped guard ring surrounding the central n^+ region, so that the breakdown voltage around the periphery is higher and avalanche is confined to the illuminated area.

The quantum efficiency is near 100% at a wavelength of about $0.75\ \mu\text{m}$ for a device having a SiO_2 - Si_3N_4 antireflection coating (Fig. 7d). Because the ratio of α_p/α_n is about 0.04, the noise factor obtained from Eq. 12 is 2.3 for $M = 10$.

10.1.6 Phototransistor

A phototransistor can have high gain through the internal bipolar-transistor action. On the other hand, the fabrication of a phototransistor is more complicated than that of a photodiode, and the inherent larger area degrades its high-frequency performance. Compared to an avalanche photodiode, it eliminates the high voltage required and high noise associated with avalanche, yet provides reasonable photocurrent gain.

A structure of bipolar phototransistor is shown in Fig. 8a, together with its circuit model in Fig. 8b. It differs from a conventional bipolar transistor in having a large base-collector junction as the light-collecting element, represented by a parallel combination of a diode and a capacitor in the model.

The phototransistor is biased in the active regime. For an n - p - n structure with a floating base, the collector is positively biased with respect to the emitter. This simply means that the collector-base junction is reversed biased and emitter-base junction is forward biased. The energy-band diagram illustrating the response to light is shown in Fig. 8c. The photogenerated holes, in the base-collector depletion region and within a distance of the diffusion length, flow to the energy maximum and are trapped in the base. This accumulation of holes or positive charges lowers the base energy (raises the potential) and allows a large flow of electrons from the emitter to the collector due to the exponential relationship between I_E and V_{BE} , i.e., $I_E \propto e^{qV_{BE}/kT}$. The result of a much larger electron current caused by a small hole current is the consequence of emitter injection efficiency γ and is the dominant gain mechanism that is common to the bipolar transistor and the phototransistor, provided that the electron transit time through the base is much shorter than the minority-carrier lifetime. The photogenerated electrons in the base-collector depletion region that are within a diffusion length distance can flow to the emitter or to the collector, depending on the location of origin. Strictly speaking, they can reduce the emitter current or enhance the collector current, but only by a very small amount since the gain is large and the total collector current or emitter current is much larger than the photocurrent.

For simplicity, we assume that light is absorbed near the base-collector junction. Since the base is open, we have $I_E = I_C$. From Fig. 8c and using the conventional bipolar transistor parameters, the total collector current is given by

$$I_C = I_{ph} + I_{CO} + \alpha_T I_{nE}, \quad (13)$$

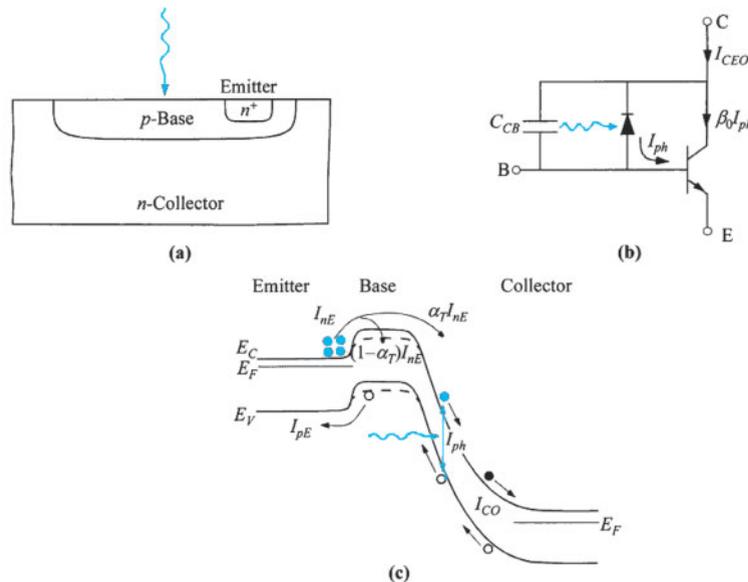


Fig. 8 (a) Schematic structure of phototransistor. (b) Equivalent circuit. (c) Energy-band diagram under bias showing different current components. Dashed lines indicate the shift of base potential (open base) under illumination.

where I_{ph} is the photocurrent, I_{CO} is the reverse leakage current of the collector-base junction, and α_T is the base transport factor. Since the base is open, the net base current is zero and

$$I_{pE} + (1 - \alpha_T)I_{nE} = I_{ph} + I_{CO}. \quad (14)$$

From Eqs. 13 and 14, and the definition of emitter efficiency γ , we have

$$I_{nE} = \gamma I_E. \quad (15)$$

Then, Eq. 15 is changed to

$$I_C = I_E = I_{CEO} = (I_{ph} + I_{CO})(\beta_o + 1) \sim \beta_o I_{ph}. \quad (16)$$

The I-V characteristics of a phototransistor under different light intensities are similar to those of the bipolar transistor, except the base incremental current is replaced with increasing light intensities. Equation 16 indicates a photocurrent gain of $(\beta_o + 1)$. In practical homojunction phototransistors, gains vary from 50 to a few hundred. The heterojunction phototransistor, whose emitter has a larger energy gap than base, can have advantages similar to that of a regular heterojunction bipolar transistor. Gains up to 10,000 can be obtained. Unfortunately, the dark current is also amplified by the same factor.

This device is particularly useful in opto-isolator applications because it offers high current-transfer ratios, i.e., the ratio of output photodetector current to the input light-source (LED or laser) current, of the order of 50% or more, as compared to a typical photodiode with a current-transfer ratio of 0.2%.

10.1.7 Heterojunction Photodiode

A heterojunction device is formed by depositing a large-bandgap semiconductor epitaxially on a smaller-bandgap semiconductor. One advantage of a heterojunction photodiode is that the quantum efficiency does not depend critically on the distance of the junction from the surface, because the large-bandgap material can be used as a window for the transmission of optical power. In addition, the heterojunction can provide unique material combinations so that the quantum efficiency and response speed can be optimized for a given optical-signal wavelength.

To obtain a heterojunction with low leakage current, the lattice constants of the two semiconductors must be closely matched. Ternary III-V compounds $\text{Al}_x\text{Ga}_{1-x}\text{As}$ epitaxially grown on GaAs can form heterojunctions with perfectly matched lattices. These heterojunctions are important for photonic devices operated in the wavelength range from 0.65 to 0.85 μm . At longer wavelengths (1 to 1.6 μm), ternary compounds such as $\text{Ga}_{0.47}\text{In}_{0.53}\text{As}$ (with $E_g = 0.75$ eV) and quaternary compounds such as $\text{Ga}_{0.27}\text{In}_{0.73}\text{As}_{0.63}\text{P}_{0.37}$ (with $E_g = 0.95$ eV) can be used. These compounds have a nearly perfect lattice match to the InP substrate. This device has superior performance to the Ge photodiode because of the direct bandgap, which gives rise to a larger absorption coefficient, so that a thinner depletion width can be used to give a higher response speed. The quantum efficiency is greater than 70% over the wavelength range from 1 to 1.6 μm , as shown in Fig. 2 (GaInAs curve).

10.1.8 Superlattice APD

As mentioned before, the APD exhibits excess noise from the random nature of the avalanche multiplication process. This avalanche noise is minimized when only the electron is involved. Figure 9a shows the energy band diagram of a staircase superlattice APD to achieve only electron avalanche multiplication.³ The energy band gap of each layer changes from a minimum E_{g1} to a maximum E_{g2} that is more than twice E_{g1} . The ΔE_C in the conduction band between two neighboring layers is larger than E_{g1} .

Under bias, as shown in Fig. 9b, the photogenerated electrons drift in the graded layer conduction band and then drift into the neighboring layer; they will have kinetic energy ΔE_C as a result of the transition and $\Delta E_C (> E_{g1})$ is sufficiently high to cause the impact ionization there. Therefore, the device does not need the high field typical of avalanche multiplication in a bulk semiconductor. The impact ionized holes experience only a small ΔE_V that is insufficient to cause impact ionization. The staircase superlattice APD exhibits low avalanche noise. However, the staircase superlattice APD is difficult to fabricate due to the graded bandgap.

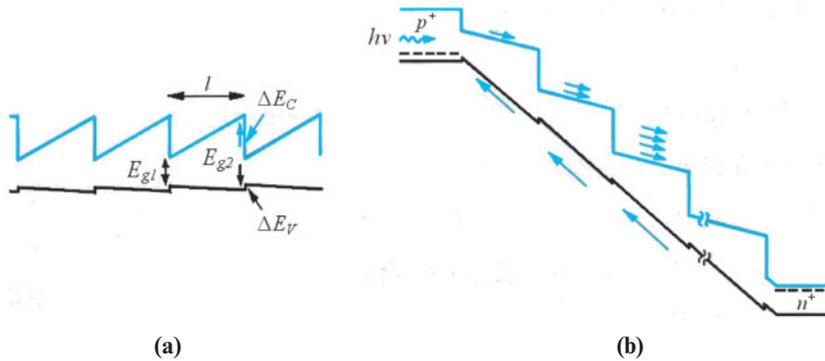


Fig. 9 Energy band diagram of a staircase superlattice APD. (a) Thermal equilibrium. (b) With biasing.³

10.1.9 Quantum-Well Infrared Photodetector

The quantum-well infrared photodetector (QWIP) is based on photoconductivity due to intersubband excitation.¹ The infrared absorption of QWIP is within the conduction band or the valence band, instead of band-to-band, in a quantum well. The three types of transitions are depicted in Fig. 10. In the bound-to-bound transition, two quantized energy states are confined and below the barrier energy. A photon excites an electron from the ground state to the first bound state and the electron subsequently tunnels out of the well. In the bound-to-continuum (or bound-to-extended) excitation, the first state above the ground state is over the barrier and excited electrons can escape the well more easily. This bound-to-continuum excitation is more promising in that it has higher absorption, broader wavelength response, lower dark current, and higher detectivity, and requires lower voltage. In the bound-to-miniband transition, a miniband is present because of the superlattice structure. QWIPs based on this have shown great promise for focal-plane array-imaging sensor system applications.

The structure of a QWIP using a GaAs/AlGaAs heterostructure is shown in Fig. 11. The quantum-well layers, in this case GaAs, have a thickness of about 5 nm and are usually doped to n -type in the 10^{17} cm^{-3} range. The barrier layers are undoped and have thickness in the range of 30-50 nm. A typical number of periods is between 20 and 50.

For quantum wells formed by direct-bandgap materials, the incident light normal to the surface has zero absorption because intersubband transitions require that the electric field of the electromagnetic wave has components normal to the quantum-well plane. This polarization selection rule demands techniques to couple light to the light-sensitive area. In Fig. 11a, a polished 45° -facet is made at the edge adjacent to the detector. Notice that the wavelength of interest is transparent to the substrate. In Fig. 11b, a grating on the top surface refracts light back to the detector. Alternatively, a grating can be made on the substrate surface to scatter the incoming light.

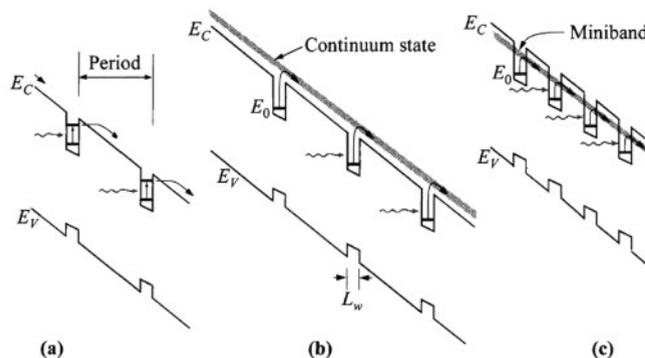


Fig. 10 Energy-band diagrams of QWIPs under bias showing three types of transition. (a) Bound-to-bound intersubband transition. (b) Bound-to-continuum transition. (c) Bound-to-miniband transition.¹

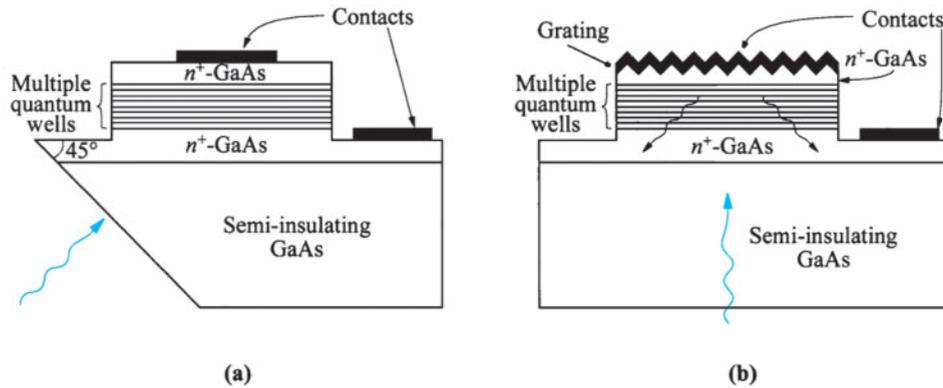


Fig. 11 Structures of GaAs/AlGaAs QWIPs showing approaches to coupling light to the heterointerface at a critical angle. (a) Light incident normal to a polished facet makes a 45° angle to the quantum well. (b) A grating is used to refract light coming from the substrate.¹

The QWIP is an attractive alternative for long-wavelength photodetectors that use HgCdTe material, which has problems of excessive tunneling of dark current and reproducibility of the precise composition required to produce the exact energy gap. Moreover, the QWIP is compatible with GaAs technology and circuits for monolithic integration. The detection wavelength range can also be tuned by the quantum-well thickness, and the long-wavelength capability can be close to 20 μm . It has high speed and fast response due to its intrinsic short carrier lifetime in the quantum wells. One difficulty with the QWIP, at least for *n*-type GaAs wells, is the detection of normal-incidence light.

► 10.2 SOLAR CELLS

Solar cells are useful for both space and terrestrial applications. Solar cells furnish the long-duration power supply for satellites. The solar cell is an important candidate for an alternative terrestrial energy source because it converts sunlight directly to electricity with good conversion efficiency, provides nearly permanent power at low operating cost, and is virtually nonpolluting.^{4,5}

10.2.1 Solar Radiation

The radiative energy output from the sun derives from a nuclear fusion reaction. Every second, about 6×10^{11} kg hydrogen is converted to helium, with a net mass loss of about 4×10^3 kg. The mass loss is converted through the Einstein relation ($E = mc^2$) to 4×10^{20} J. This energy is emitted primarily as electromagnetic radiation in the ultraviolet to infrared region (0.2 to 3 μm). The total mass of the sun is now about 2×10^{30} kg, and a reasonably stable life with a nearly constant radiative-energy output of over 10 billion (10^{10}) years is projected.

The intensity of solar radiation outside the earth's atmosphere, at the average distance of its orbit around the sun, is defined as the solar constant and has a value of 1367 W/m^2 . Terrestrially, the sunlight is attenuated by clouds and by atmospheric scattering and absorption. The attenuation depends primarily on the length of the light's path through the atmosphere, or the mass of air through which it passes. This "air mass" is defined as $1/\cos \phi$, where ϕ is the angle between the vertical and the sun's position.

► EXAMPLE 4

The air mass can most easily be estimated from the length of the shadow, s , of a vertical structure of height h , as $\sqrt{1 + (s/h)^2}$. If $s = 1.118 \text{ m}$ and $h = 1.00 \text{ m}$, find the air mass.

SOLUTION

$$\sqrt{1 + (1.118/1.0)^2} = \sqrt{2.25} = 1.5.$$

We have an air mass 1.5 (AM 1.5). The corresponding $\cos \phi$ is $1/1.5 = 0.667$ and the angle ϕ between the vertical and the sun's position is $\cos^{-1}(0.667) = 48^\circ$. The maximum sunlight intensity occurs when the sun is straight overhead (i.e., AM 1.0 with $\phi = 0^\circ$).

Figure 12 shows two curves related to solar spectral irradiance (power per unit area per unit wavelength).⁶ The upper curve, which represents the solar spectrum outside the Earth's atmosphere, is the air mass zero condition (AM0). The AM0 spectrum is relevant for satellite and space vehicle applications. Terrestrial solar-cell performance is specified with reference to the air mass 1.5 (AM 1.5) spectrum. This spectrum represents the sunlight at the Earth's surface when the sun is at an angle of 48° from the vertical. At this angle the incident power is about 963 W/m^2 .

10.2.2 *p-n* Junction Solar Cell

A schematic representation of a *p-n* junction solar cell is shown in Fig. 13. It consists of a shallow *p-n* junction formed on the surface, a front ohmic contact stripe and fingers, a back ohmic contact that covers the entire back surface, and an antireflection coating on the front surface. The surface reflection of the incident light from air ($\bar{n} = 1$) into semiconductor silicon ($\bar{n} = 3.5$) is about 0.31. This means that 31% of incident light is reflected and is not available for conversion to electrical energy in a silicon solar cell.

When the cell is exposed to the solar spectrum, a photon that has an energy less than the bandgap E_g makes no contribution to the cell output. A photon that has energy greater than E_g contributes an energy E_g to the cell output. Energy greater than E_g is wasted as heat. When EHPs are created in the depletion layer, they are separated by the built-in electric field. Hence, the potential difference is limited by the built-in voltage, which is in turn determined by the energy gap. On the other hand, only photons with energies larger than the bandgap are absorbed in a semiconductor, and hence the light-generated current decreases with the increase in energy gap due to the limited solar spectrum.

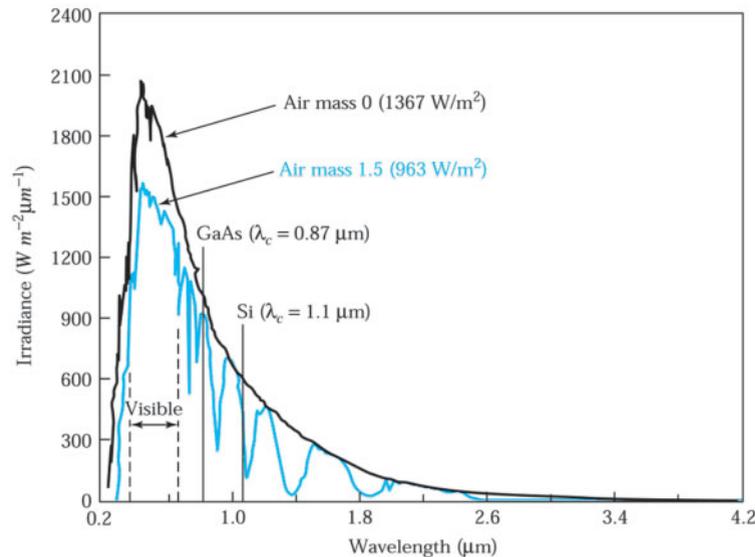


Fig. 12 Solar spectral irradiance⁶ at air mass 0 and air mass 1.5 and the cutoff wavelength of GaAs and Si.

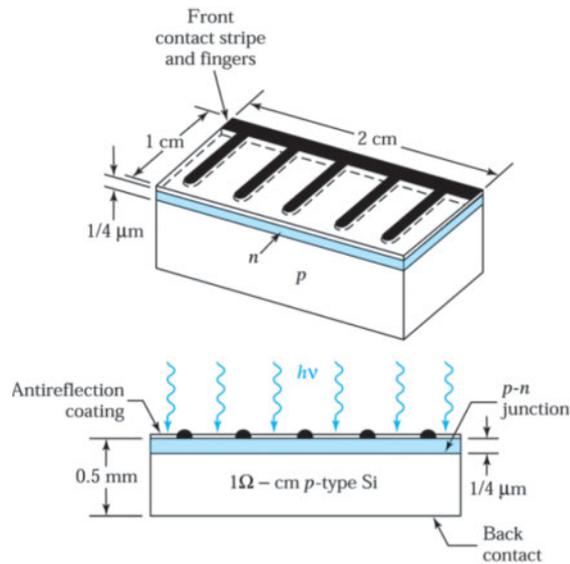


Fig. 13 Schematic representation of a silicon p - n junction solar cell.⁴

To derive the conversion efficiency, we consider the energy band diagram of a p - n junction, shown in Fig. 14a, under solar radiation. We can see that V_{OC} depends on the light intensity. The efficiency does not depend critically on the bandgap. Semiconductors with bandgaps between 1 and 2 eV can all be considered solar cell materials. The equivalent circuit is shown in Fig. 14b, where a constant-current source is in parallel with the junction. The source I_L results from the excitation of excess carriers by solar radiation, I_s is the diode saturation current, and R_L is the load resistance.

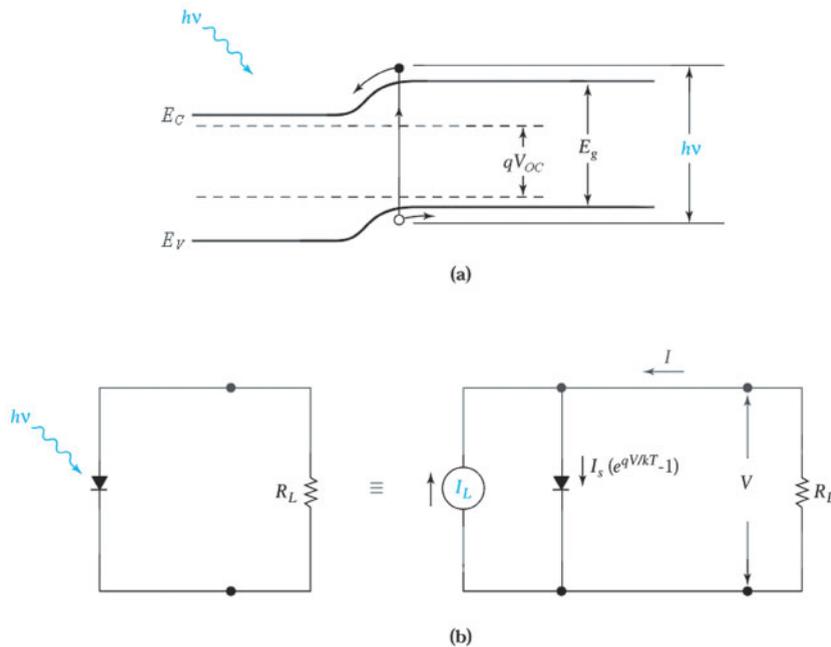


Fig. 14 (a) Energy band diagram of a p - n junction solar cell under solar irradiation. (b) Idealized equivalent circuit of a solar cell.

The ideal I - V characteristics of such a device are given by

$$I = I_s(e^{qV/kT} - 1) - I_L, \quad (17)$$

and

$$J_s = \frac{I_s}{A} = qN_C N_V \left(\frac{1}{N_A} \sqrt{\frac{D_n}{\tau_n}} + \frac{1}{N_D} \sqrt{\frac{D_p}{\tau_p}} \right) \cdot e^{-E_g/kT}, \quad (17a)$$

where A is the device area. A plot of Eq. 17 is given in Fig. 15a for $I_L = 100$ mA, $I_s = 1$ nA, cell area $A = 4$ cm², and $T = 300$ K. The curve passes through the fourth quadrant, and therefore power can be extracted from the device. The I - V curve is more generally represented by Fig. 15b, which is an inversion of Fig. 15a about the voltage axis. A load R_L is connected to the solar cell as shown in Fig. 14b. The current through the R_L is in the opposite direction to the conventional current flow. Thus,

$$I = -V/R_L \quad (18)$$

This current and the current in the circuit must satisfy both the I - V characteristics of the solar cell Eq. 17 and that of the load Eq. 18 simultaneously. The load line with slope $-1/R_L$ is shown in Fig. 15a. The intersection point is the operating point at which the load and the solar cell have the same current and voltage. By choosing a proper load, close to 80% of the product $I_{sc} V_{oc}$ can be extracted, where I_{sc} is the short-circuit current and V_{oc} is the open-circuit voltage of the cell; the shaded area in the figure is the maximum-power rectangle. Also defined in Fig. 15b are the quantities I_m and V_m that correspond to the current and voltage, respectively, for the maximum power output $P_m (= I_m \times V_m)$.

From Eq. 17 we obtain for the open-circuit voltage ($I = 0$)

$$V_{oc} = \frac{kT}{q} \ln \left(\frac{I_L}{I_s} + 1 \right) \cong \frac{kT}{q} \ln \left(\frac{I_L}{I_s} \right). \quad (19)$$

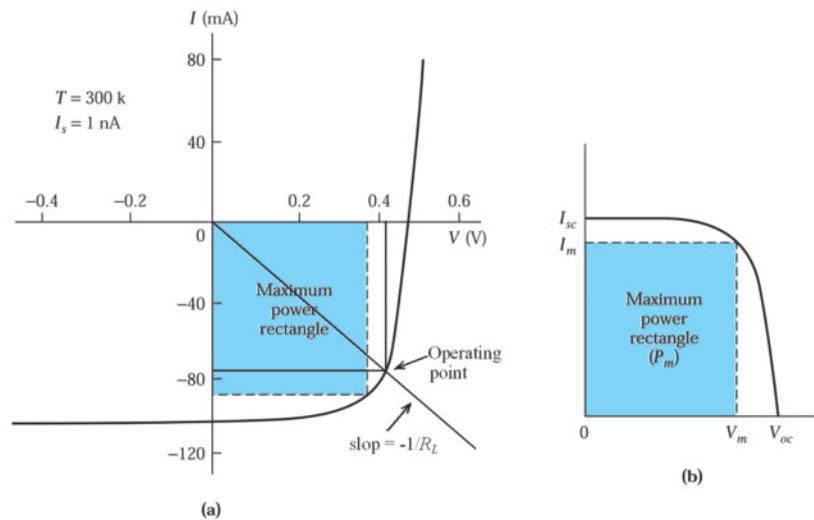


Fig. 15 (a) Current-voltage characteristics of a solar cell under illumination. (b) Inversion of (a) about the voltage axis.

Hence, for a given I_L , V_{OC} increases logarithmically with decreasing saturation current I_s . The output power is given by

$$P = IV = I_s V (e^{qV/kT} - 1) - I_L V. \quad (20)$$

The condition for maximum power is obtained when $dP/dV = 0$, or

$$V_m = \frac{kT}{q} \ln \left[\frac{1 + (I_L / I_s)}{1 + (qV_m / kT)} \right] \cong V_{OC} - \frac{kT}{q} \ln \left(1 + \frac{qV_m}{kT} \right), \quad (21a)$$

$$I_m = I_s \left(\frac{qV_m}{kT} \right) e^{qV_m/kT} \cong I_L \left(1 - \frac{1}{qV_m / kT} \right). \quad (21b)$$

The maximum output power P_m is then

$$P_m = I_m V_m \cong I_L \left[V_{OC} - \frac{kT}{q} \ln \left(1 + \frac{qV_m}{kT} \right) - \frac{kT}{q} \right]. \quad (22)$$

► EXAMPLE 5

Calculate the open-circuit voltage and the output power at a voltage of 0.35 V for the solar cell shown in Fig. 15a.

SOLUTION From Eq. 19, we have

$$V_{OC} = (0.026 \text{ V}) \ln \left(\frac{100 \times 10^{-3} \text{ A}}{1 \times 10^{-9} \text{ A}} \right) = 0.48 \text{ V}$$

The output power at 0.35 V is given by Eq. 20 (note that I_s and I_L are reverse current so we need negative signs for them):

$$P = (-10^{-9} \text{ A}) \cdot (0.35 \text{ V}) (e^{0.35/0.026} - 1) - (-0.1 \text{ A}) \cdot (0.35 \text{ V}) = 3.48 \times 10^{-2} \text{ W}. \quad \blacktriangleleft$$

10.2.3 Conversion Efficiency

Ideal efficiency

The power conversion efficiency of a solar cell is given by

$$\eta = \frac{I_m V_m}{P_{in}} = \frac{I_L \left[V_{OC} - \frac{kT}{q} \ln \left(1 + \frac{qV_m}{kT} \right) - \frac{kT}{q} \right]}{P_{in}} \quad (23)$$

or

$$\eta = \frac{FF \cdot I_{sc} V_{oc}}{P_{in}}, \quad (23a)$$

where P_{in} is the incident power and FF is the fill factor defined as

$$FF \equiv \frac{I_m V_m}{I_{sc} V_{oc}} \cong 1 - \frac{kT}{qV_{oc}} \ln\left(1 + \frac{qV_m}{kT}\right) \approx \frac{kT}{qV_{oc}}, \quad (24)$$

assuming $I_{sc} \cong I_L$. The fill factor is the ratio of the maximum power rectangle (Fig. 15b) to the rectangle of $I_{sc} \times V_{oc}$. In practice, a good fill factor is around 0.8. To maximize the efficiency, we should maximize all three items in the numerator of Eq. 23a.

The ideal efficiency can be obtained from the ideal I - V characteristics defined by Eq. 17. For a given semiconductor, the saturation current density is obtained from Eq. 17a. For a given air mass condition (e.g., AM 1.5), the short-circuit current I_L is the product of q and the number of the available photons with energy $h\nu \geq E_g$ in the solar spectrum. Once I_s and I_L are known, the output power P and the maximum power P_m can be obtained from Eqs. 20 through 22. The input power P_{in} is the integration of all the photons in the solar spectrum (Fig. 12). Under AM 1.5 condition, the efficiency P_m/P_{in} has a broad maximum^{5,7} of about 29% and does not depend critically on E_g . Therefore, semiconductors with bandgap between 1 and 2 eV can all be considered as solar cell materials. Many factors degrade the ideal efficiency, so that efficiencies actually achieved are lower. The ideal peak efficiency is 31% for one sun and 37% for 1000 suns.^{1,7}

Spectrum Splitting

The simplest way to improve the efficiency is by spectrum splitting. By splitting sunlight into narrow wavelength bands and directing each band to a cell that has a bandgap optimally chosen to convert just this band, as shown in Fig. 16a, efficiency above 60% is in principle possible.⁸ Fortunately, simply stacking cells on top of one another with the highest bandgap cell uppermost, as in Fig. 16b, automatically achieves an identical spectral-splitting effect, making this “tandem” cell approach a reasonably practical way of increasing cell efficiency.

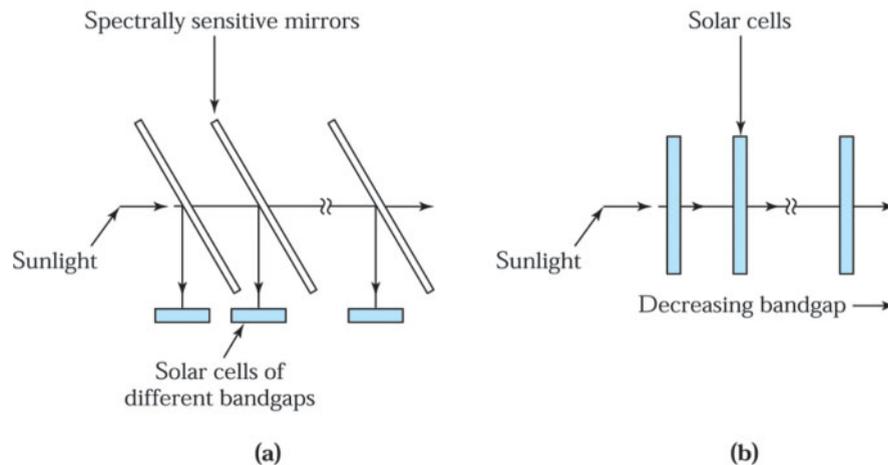


Fig. 16 Multigap cell concepts. (a) Spectrum-splitting approach. (b) Tandem-cell approach.⁸

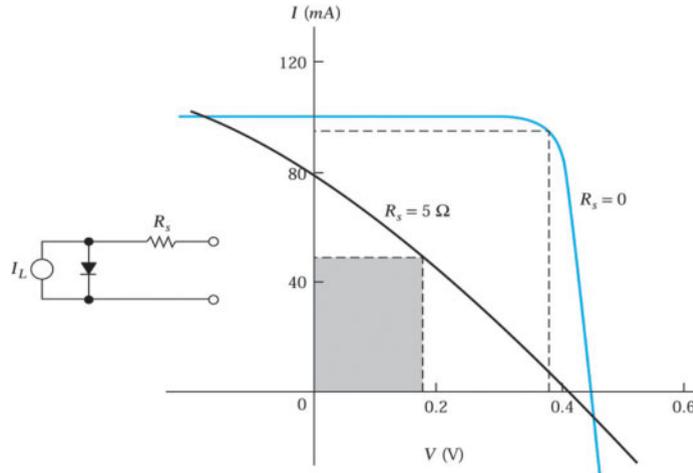


Fig. 17 Current-voltage characteristics and the equivalent circuit of solar cells that have resistances.

Series Resistance and Recombination Current

Many factors degrade the ideal efficiency. One of the major factors is the series resistance R_s from the ohmic loss in the front surface. As shown in Fig. 13, the photogenerated electrons traverse the n -layer to the finger electrodes and introduce an effective series resistance. If the finger electrodes are thin, the series resistance will be further increased. There is also a series resistance in the p -region, but it is generally small due to the bulky volume. On the other hand, a shunt resistance will also be present because a fraction (usually small) of the photogenerated carriers can flow through the crystal surface (or through grain boundaries in polycrystalline devices) instead of through the external load. Typically the shunt resistance is less important than the series resistance. The equivalent circuit is shown in Fig. 17. From the ideal diode current given by Eq. 17, the I - V characteristics are found to be

$$\ln\left(\frac{I+I_L}{I_s}+1\right) = \frac{q}{kT}(V - IR_s). \quad (25)$$

A plot of this equation is shown in Fig. 17, with $R_s = 0$ and 5Ω and where the other parameters I_s , I_L , and T are the same as those in Fig. 15. It can be seen that a series resistance of only 5Ω reduces the available power to less than 30% of the maximum power with $R_s = 0$. The output current and output power are

$$I = I_s \left\{ \exp\left[\frac{q(V - IR_s)}{kT}\right] - 1 \right\} - I_L, \quad (26)$$

$$P = I \left[\frac{kT}{q} \ln\left(\frac{I+I_L}{I_s}+1\right) + IR_s \right]. \quad (27)$$

The series resistance depends on the junction depth, the impurity concentrations of p -type and n -type regions, and the arrangement of the front-surface ohmic contacts. For a typical silicon solar cell with the geometry shown in Fig. 13, the series resistance is about 0.7Ω for n^+ - p cells and 0.4Ω for p^+ - n cells. The difference in resistance is mainly the result of the lower resistivity in n -type substrates.

Another factor is the recombination current in the depletion region. For single-level centers, the recombination current can be expressed as

$$I_{rec} = I_s' \left[\exp\left(\frac{qV}{2kT}\right) - 1 \right], \quad (28)$$

and

$$\frac{I_s'}{A} = \frac{qn_i W}{\sqrt{\tau_p \tau_n}}, \quad (28a)$$

where I_s' is the saturation current. The energy conversion equation can be put into closed form to yield equations similar to Eqs. 19 through 22, with the exception that I_s is replaced by I_s' and the exponential factor is divided by 2. The efficiency for the recombination current case is found to be much less than the ideal current due to the degradation of both V_{oc} and the fill factor. For silicon solar cells at 300 K, the recombination current can cause a 25% reduction in efficiency.

► 10.3 SILICON AND COMPOUND-SEMICONDUCTOR SOLAR CELLS

The main requirements for solar cells are high efficiency, low cost, and good reliability. Many solar-cell configurations have been proposed and demonstrated with impressive results. However, for solar cells to supply a significant portion of world energy, more challenges are still ahead. Nevertheless, we believe that the goal is achievable. We consider a few key solar-cell designs and their performances. In general, there are two categories of solar cells: wafer-based and thin-film solar cells.

10.3.1 Wafer-Based Solar Cells

Silicon is the most important semiconductor for solar cells. It is nontoxic and is second only to oxygen in prevalence in the earth's crust. Therefore, silicon poses minimal environmental or resource-depletion risks if used on a large scale. It also has a well established technological base because of its use in microelectronics.

III-V compound semiconductors and their alloy systems provide wide choices of bandgaps with closely matched lattice constants. These compounds are ideal for producing tandem solar cells. For example, AlGaAs/GaAs, GaInP/GaAs, and GaInAs/InP material systems have been developed for solar cells in satellite and space vehicle applications.

Silicon PERL Cell

Usually, short-circuit current losses come from metal-finger coverage of the top surface, top-surface reflection loss, and imperfect light trapping in the cell. The voltage losses arise from finite surface and bulk recombination. The fill factor losses come not only from ohmic series resistance loss within the cell, but also from the same factors producing the open-circuit voltage loss. The silicon passivated emitter and rear locally-diffused (PERL) cell⁹ shown in Fig. 18a is a solar cell design taking all those loss factors into account.

The cell has inverted pyramids on the top that are formed by using anisotropic etches to expose the slowly etching (111) crystallographic planes. The pyramids reduce reflections of light incident on the top surface, since incident light perpendicular to the cell will strike one of the inclined (111) planes obliquely and will be refracted obliquely into the cell. This enhanced light trapping reduces the short-circuit current loss.

The cell is characterized by the use of a thin, thermally grown oxide to “passivate” (reduce the electronic activity of) the top surface of Si wafer for a junction diffusion. Then, a shallow, low-sheet-resistivity phosphorus diffusion n -layer is formed. The oxide passivation of the cell surfaces can improve the open-circuit voltage. It can also function as an antireflection coating with refractive index $\bar{n} = 1.46$ to further reduce the total reflection. The rear locally diffused region is formed in the area of the rear point contact.

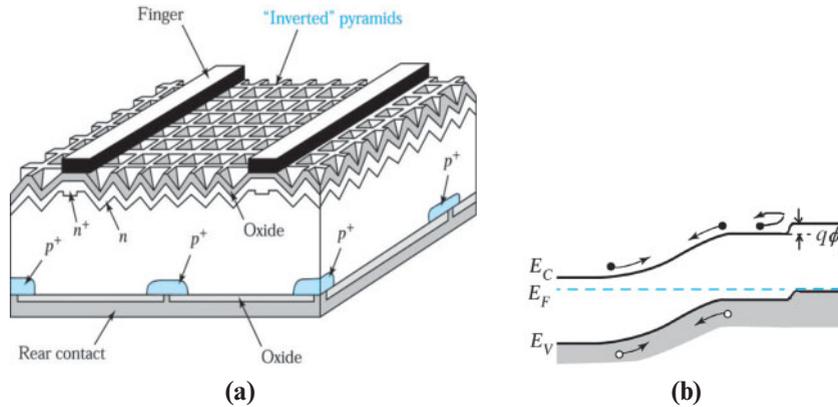


Fig. 18 (a) Passivated emitter rear locally diffused (PERL) cell.⁹ (b) Energy-band diagram for the back-surface field.

The incorporation of a heavily doped layer under the back contact, a so-called “back-surface field,” is shown in Fig. 18b. The potential energy $q\phi_p$ provides a minority carrier-reflecting region between this contact and the substrate. The back surface field also results in a very small recombination velocity at the back. Therefore, the short-circuit current will increase. The open-circuit voltage is also increased due to the increased short-circuit current. It also reduces the contact resistance and improves the fill factor. The rear contact is separated from the silicon by an intervening oxide layer. This gives much better rear reflection than an aluminum layer. To date, the PERL cell shows the highest conversion efficiency of 24.7%.

III-V Compound Tandem Solar Cell

A major factor limiting conversion efficiency in single bandgap cells to 31% is that the absorbed photon energy above the semiconductor band gap is lost as heat. The main approach to reducing this efficiency loss is to use tandem $p-n$ junctions in which higher-bandgap semiconductors and lower-bandgap semiconductors are connected together with a p^+-n^+ tunneling diode. Higher-energy photons are absorbed in the higher-bandgap semiconductors and lower-energy photons in the lower-bandgap semiconductors with band gaps better matched to the solar spectrum, and the overall heat loss is reduced. Stacking dozens of different cells together can theoretically increase efficiency to 68%. But this results in technical problems such as strain damages to the crystal layers. The most efficient multi-junction solar cell is one that has three cells.

Figure 19 shows the structure of a monolithic tandem solar cell.¹ A p -type germanium, which has a lattice constant very close to that of GaAs and $\text{Ga}_{0.51}\text{In}_{0.49}\text{P}$, is used as the substrate. The top junction is the GaInP junction ($E_g = 1.9$ eV), which can absorb photons with energy $h\nu > 1.9$ eV. The bottom junction is the GaAs $p-n$ junction ($E_g = 1.42$ eV), which can absorb photons with energy 1.9 eV $> h\nu > 1.42$ eV. A tunneling p^+-n^+ GaAs junction is placed between the top and bottom junctions to connect the cells. A p -AlGaInP layer is grown below the top junction to form a high-low junction p -AlGaInP/ p -GaInP, and a p -GaInP layer is grown below the bottom junction to form a high-low junction p -GaInP/ p -GaAs. They also function as a “back surface field,” as mentioned above. The potential energy barrier $q\phi_p$ for the back surface field can be higher for heterojunctions than that for $p-p^+$ homojunctions and drives minority carriers (electrons) back in the lower bandgap region of the high-low junction. There is a window at the top of each cell. A narrow layer of a wide bandgap semiconductor serves as a window, n -AlInP for the top cell and n -GaInP for the bottom cell, for the sunlight that reaches the narrow-gap semiconductor with little loss. These layers can passivate the surface defects normally present in a homojunction cell, thereby overcoming the surface recombination and improving cell efficiency. The window layer is typically very heavily doped.

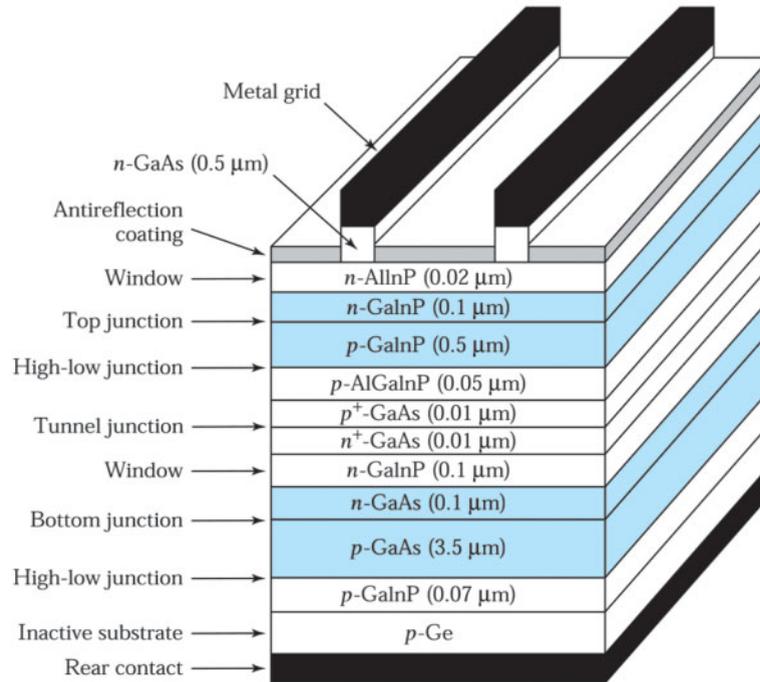


Fig. 19 Monolithic tandem solar cell.¹

It has a higher built-in voltage and hence a higher open-circuit voltage, and a higher cell efficiency. High doping also reduces the parasitic series resistance. Similar InGaP/GaAs/InGaAs 3-junction cells grown on a Ge substrate show a higher efficiency. Tandem solar cells with efficiency as high as 40% have been obtained.¹⁰

10.3.2 Thin-Film Solar Cells

The biggest problem with the conventional Si solar cell is cost. It requires a relatively thick layer of single crystalline silicon in order to have reasonable photon capture rates, and such silicon is an expensive commodity. The thin-film solar cell can provide a lower-cost alternate approach.

Amorphous Si Solar Cell

Amorphous silicon (*a*-Si) thin films can be deposited directly on low-cost large-area substrates. In amorphous silicon, the distribution of bond lengths and bond angles disturb the long-range order of the crystalline silicon lattice and change the optical and electronic properties. The optical energy gap increases from 1.12 eV of single crystalline silicon to about 1.7 eV. Due to internal scattering, the apparent optical absorption is nearly an order of magnitude higher than the crystalline material.

The basic cell structure for a series interconnected *a*-Si solar cells is shown¹¹ in Fig. 20. A layer of SiO₂ followed by a transparent conducting layer of a large bandgap, degenerately doped semiconductor such as SnO₂ is deposited onto a glass substrate and patterned using a laser. The substrate is then coated by a *p-i-n* junction stack of amorphous silicon by the decomposition of silane in a radio-frequency plasma-discharge system. After deposition, the *a*-Si layers are patterned by a laser system. A layer of aluminum is sputtered onto the rest of the silicon and this layer is also patterned by laser. This technique forms a series of interconnected cells, as shown in Fig. 20. The cell has the lowest manufacturing cost but a modest efficiency of 6%.

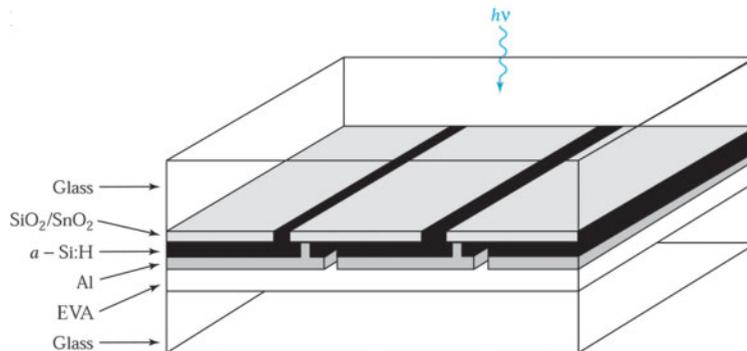


Fig. 20 Series-interconnected a-Si solar cells deposited on a glass substrate with a rear glass cover bonded using ethylene vinyl acetate (EVA).¹¹

The amorphous silicon obtained by this process is incorporated with a fairly large concentration of hydrogen. The hydrogen atoms tie up dangling silicon bonds and decrease the density of localized states in the energy gap. These localized states play a dominant role in determining the carrier transport properties of amorphous silicon. The typical deposition temperature is below 300 °C, otherwise no hydrogen is incorporated in the film.

Due to the low carrier mobilities, the collection of photogenerated carriers has to be supported by an internal electrical field. To create a high field in the intrinsic layer of the *p-i-n* structure the cells have to be thin, of the order of a few hundred nanometers. For the *p-i-n* structure, the *p*- and *n*-doped layers are generally kept very thin (< 50 nm), since material quality decreases significantly as the doping level increases, and hence very few of the carriers generated in these layers contribute to photocurrent. However, these doped layers do establish an electric field in the better-quality *i*-layer (~ 0.5 μm thick), which aids the collection of carriers generated in this region.

In larger outdoor “power” modules, the beneficial effect of the hydrogen upon the amorphous-Si properties deteriorates under illumination. A steady drop of output efficiency occurs over the first few months. The stability problem is caused by the so called “*Staebler-Wronski*” degradation—the illumination by light with photon energies larger than the energy gap leads to new light-induced defect states. After that, the output stabilizes. Amorphous-Si-based modules are generally rated by manufacturers in terms of such “stabilized” output.

An improvement in efficiency can be achieved by utilizing tandem cells. High-quality *a*-Si:Ge:H alloys can be used as the narrow bandgap material. The bandgap of *a*-Si incorporated with Ge is reduced to about 1.5 eV. Therefore, we can fabricate higher efficiency *a*-Si:H/*a*-Si:Ge:H tandem cells with better collection of the red portion of the solar spectrum. A stabilized efficiency around 8% for large-area modules was obtained with these cells. A stabilized efficiency above 13% was obtained using a triple junction with the top cell consisting of a layer of *a*-Si:H and the bottom two cells having increased thicknesses and containing increasing percentage of germanium.^{5,12} But the corresponding process gas GeH₄ contributes substantially to the cost of the module.

One promising microcrystalline tandem solar cell with much higher efficiency (14.5%) than the amorphous type has been developed.^{13,14} The structure shown in Fig. 21a consists of a microcrystalline bottom cell (*μc*-Si:H) and a conventional amorphous top cell in tandem. The optical energy gap of *μc*-Si:H is somewhere around 1 eV, which is close to that of crystalline Si and very different from that of *a*-Si:H (1.7 eV).

The short-wavelength light is absorbed by the top amorphous cell and long-wavelength light is absorbed by the bottom microcrystalline cell. The spectral sensitivity of the microcrystalline tandem solar cell in Fig. 21b shows higher efficiency because the microcrystalline cell absorbs the long-wavelength light that cannot be absorbed by the amorphous silicon. Compared with *a*-Si:H/*a*-Si:Ge:H tandem cells, the spectral response of the *a*-Si:H/*μc*-Si:H

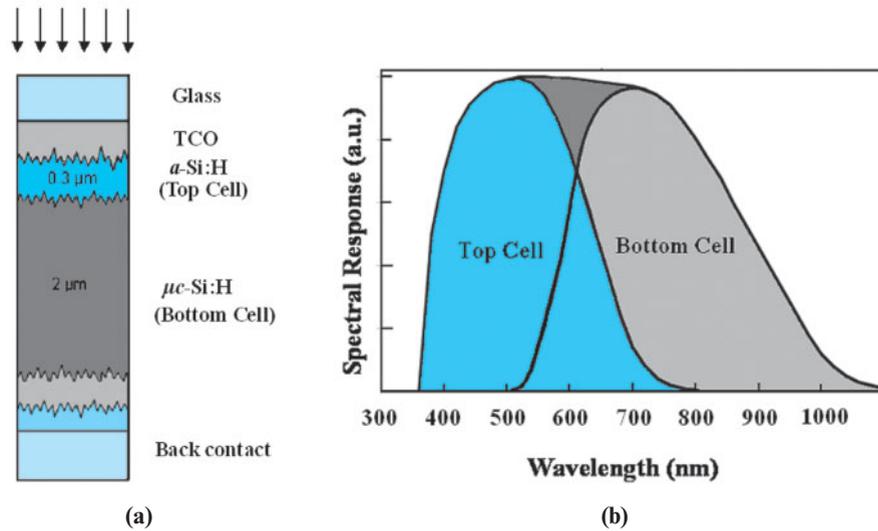


Fig. 21 (a) Schematic structure and (b) typical spectral response of a microcrystalline/ amorphous tandem cell.¹³

tandem cell is strongly extended towards longer wavelengths. Because the microcrystalline Si has a lower optical-absorption coefficient than the amorphous type, the thickness of the *i*-layer of the microcrystalline solar cell needs to be much greater than the amorphous solar cell.

CIGS Solar Cell

In 1974, Bell Telephone Laboratories reported the first copper indium diselenide (CuInSe_2) solar cell with a conversion efficiency of 6%. In 1982, the CdS/CuInSe_2 solar cell with a conversion efficiency of 10% was developed. With indium partially replaced by gallium in CuInSe_2 to form the copper indium gallium diselenide (CIGS), the new material has larger optical bandgap than pure CIS, thus increasing the open-circuit voltage. The conversion efficiency of CdS/Cu(In,Ga)Se_2 (CIGS) was raised to 15% in 1993, to 17.7% in 1996, to 19.2% in 2003,^{5,15} and to 19.9% in 2008.¹⁶

The CIS is a direct bandgap semiconductor material and its absorption coefficient is higher over a broader wavelength range than other semiconductors, as shown in Fig. 22a.¹⁷ The bandgap of CIGS can vary continuously from about 1.0 eV (for CuInSe_2) to about 1.7 eV (for CuGaSe_2). A typical structure of CIGS solar cell is shown in Fig. 22b. The soda-lime glass [the most prevalent type of glass prepared with sodium carbonate (soda), limestone, etc.] was used as the substrate. Na ions in soda-lime glass will diffuse through Mo into CIGS during growth, and the grain of polycrystalline CIGS can grow larger with fewer defects. Sodium not only improves crystallization of the film but also increases conductivity due to the sodium incorporated at grain boundaries or defects. The mechanism is still not clear. Mo with high reflectivity and low resistivity forms good ohmic contact with CIGS. *P*-type CIGS absorbs most of the light and has been deposited using various methods, including co-evaporation, reactive sputtering sublimation, chemical bath deposition, laser evaporation, and spray pyrolysis. The *p-n* heterojunction is formed by depositing a very thin *n*-type CdS and an *n*-type transparent conducting oxide ZnO (ZnO:Al). CdS is used to modify the CIGS sensitive surface and lower the band discontinuity between ZnO and CIGS. ZnS may replace CdS due to environmental concerns. The direct deposition of ZnO on CdS will induce local defects (such as pin holes) and local fluctuations in CIGS properties (e.g., bandgap). An intrinsic ZnO (*i*-ZnO) buffer layer would decrease these problems. MgF_2 is used as an anti-reflection coating. CIGS-based solar cells are presently one of the best candidates for a new generation of large-scale, low-cost thin-film photovoltaic systems.

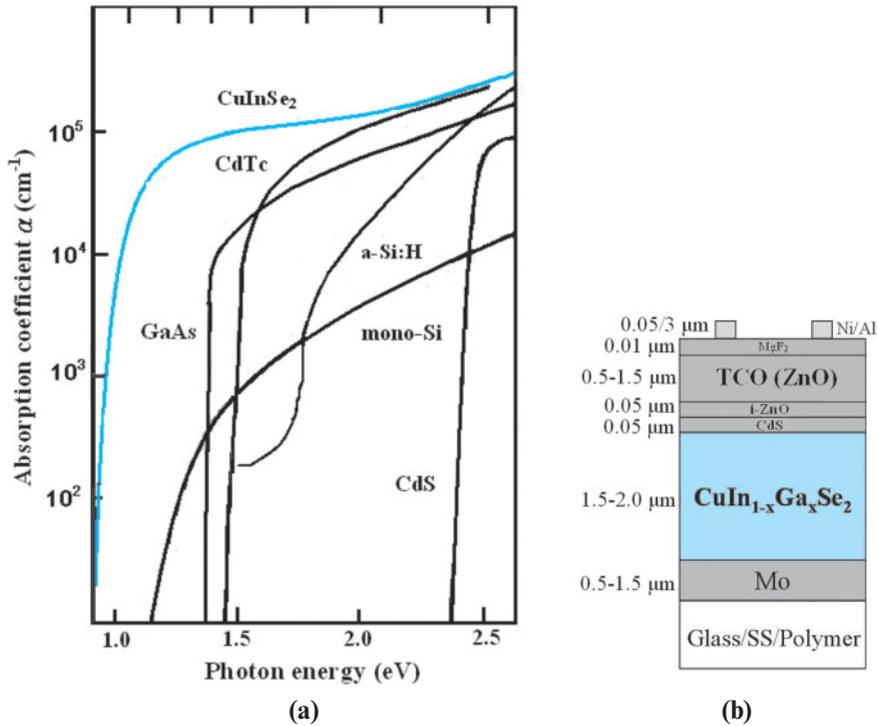


Fig. 22 (a) Optical absorption coefficient of CuInSe_2 . (b) A typical structure of CIGS solar cell.

► 10.4 THIRD-GENERATION SOLAR CELLS

The third generation photovoltaic cell is a range of novel alternatives to “first generation” (silicon single-crystal *p-n* junction or wafer solar cells) and “second generation” (low-cost, but low-efficiency thin-film) cells. Research and development in this area generally aim to provide higher efficiency and lower cost per watt of electricity generated.¹⁸

Dye-sensitized Solar Cells

Dye-sensitized solar cells (DSSCs) are currently the most efficient third-generation solar technology available and are ready for mass production.¹⁹ The cell in Fig. 23a has a layer of transparent conductive oxide (TCO) [usually fluorine-doped tin oxide ($\text{SnO}_2:\text{F}$)] deposited on glass used as anode. On the conductive plate is a layer of titanium dioxide (TiO_2), formed into a highly porous 3-D structure with an extremely high surface area for holding large numbers of dye molecules. The plate is then immersed in a mixture of a photosensitive ruthenium-polypyridine dye solution. The dye molecules are quite small (nanometer sized). In order to capture a reasonable amount of the incoming light, the layer of dye molecules covalently bonded on highly porous 3-D nano-structured TiO_2 surface needs to be fairly thick. A separate backing is made with a thin layer of the iodide/iodine electrolyte spread over a conductive platinum sheet.

The bulk of the semiconductor (TiO_2) is used solely for charge transport; the photoelectrons are provided from a separate photosensitive dye. Charge separation occurs at the surfaces between the dye, semiconductor, and electrolyte. Photons with enough energy will create an excited state of the dye, as shown in Fig. 23b. An excited electron in the conduction band has a probability to go back to valence band of the dye as the loss path 1. The excited electron can be injected directly into the conduction band of the TiO_2 , and from there it moves by diffusion to the anode. Meanwhile, the dye molecule strips one electron from iodide in electrolyte, oxidizing it into triiodide.

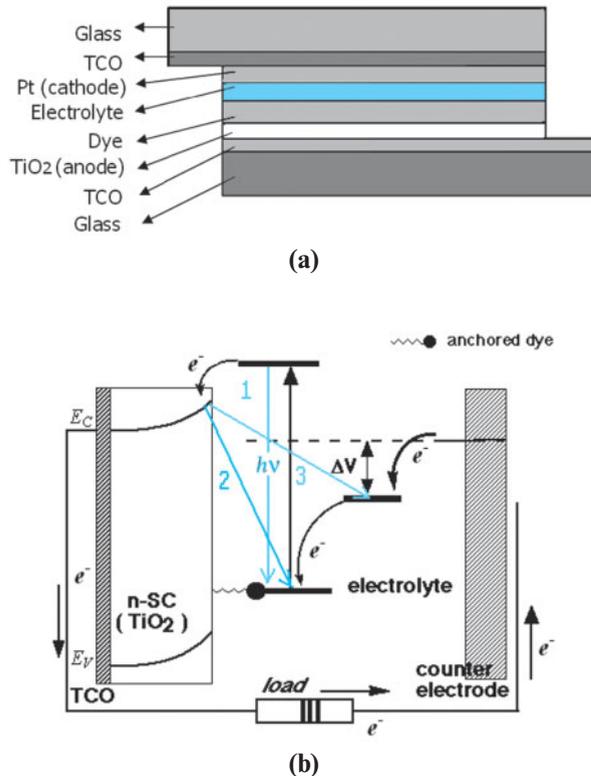


Fig. 23 (a) DSSC cell structure. (b) Energy band diagram and main carrier losses.

This reaction occurs quite quickly compared to the time for the injected electron to recombine with the oxidized dye molecule, which is the loss path 2 shown in Fig. 23b. The triiodide then recovers its missing electron by diffusing to the counter-electrode, which reintroduces the electrons after flowing through the external circuit. The third loss is from the recombination of injected electron with the electrolyte (path 3).

Due to the porosity of the TiO₂ nanostructure, there is a very high chance that a photon will be absorbed. The dye is highly efficient in converting photons into electrons, but only those electrons with enough energy can cross the TiO₂ bandgap and result in photocurrent. In addition, the electrolyte limits the speed at which the dye molecules can regain their electrons and become available for photoexcitation again. These factors limit the photocurrent generated by a DSSC. The bandgap is slightly larger than silicon, which means that fewer of the photons in sunlight can be used for carrier generation. The maximum voltage generated, in theory, is simply the difference between the Fermi level of the TiO₂ and the redox potential of the electrolyte, about 0.7 V (V_{oc}). DSSCs offer slightly higher V_{oc} than the silicon solar cell (about 0.6 V). The fill factor is about 70%, and the quantum efficiency is about 11%.²⁰

Organic Solar Cells

Carrier mobilities are very low because their transport processes are dominated by carrier hopping in organic semiconductors, as mentioned in Sec. 9.3.2 of Chapter 9, and therefore the thicknesses of organic active layers in organic solar cells are limited to a few hundred nanometers for lower series resistance. However, organic semiconductors show strong absorption in UV and visible regions and the penetration depth of the incident light is typically 80-200 nm. Thus, only a 100 nm thick organic active layer is sufficient for effective absorption. Currently, the power conversion efficiency is only 5.7%,²¹ but organic solar cells attract high interest due to their large-area, low-cost potential.

Due to electrostatic interactions, the EHP upon absorption of a photon of sufficient energy forms a tightly bound state exciton, whose binding energy is expected to be in a range of 200 ~ 500 meV. The exciton binding energy is roughly one order of magnitude larger than that for inorganic semiconductors like Si, where photoexcitations typically lead directly to free carriers at room temperature. In general, only 10% of the excitons dissociate into free carriers, while the remaining excitons decay via radiative or nonradiative recombination pathways after a short time. Thus, the energy efficiencies of single-layer polymer solar cells typically remain below 0.1%.

Solar cells with a heterojunction between donor and acceptor molecules can efficiently dissociate photogenerated excitons into free carriers at the interface and exhibit superior performances. After photoexcitation of an electron from the HOMO to the LUMO shown in Fig. 24, the electron can jump from the LUMO of the donor (with the higher LUMO) to the LUMO of the acceptor if the potential difference $\Delta\Phi$ between the ionization potential of the donor and the electron affinity of the acceptor is larger than the exciton binding energy. However, this process of so-called photo-induced charge transfer can lead to free charges only if the hole remains on the donor with the higher HOMO level. Moreover, the space between donor and acceptor should be in the range of the exciton diffusion length for efficient transfer and dissociation. A heterojunction can be prepared with donor and acceptor bilayers shown in Fig. 25a. This bilayer geometry guarantees directional photoinduced charge transfer across the interface, and the recombination losses are reduced. However, the interfacial area and thus the exciton dissociation efficiency are limited. Higher interfacial area and thus the improved exciton dissociation efficiency can be achieved if a mixture layer contains both electron donor and electron acceptor (so-called bulk heterojunctions) shown in Fig. 25b, but needs a percolating pathway for the separated charge carriers to reach their corresponding electrodes. Both approaches can be carried out either by sublimation of small molecules or by spin-coating of polymers.

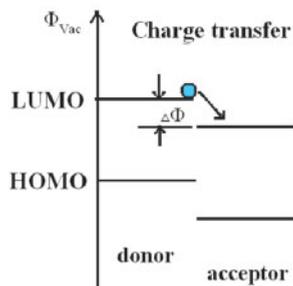


Fig. 24 Heterojunction between donor and acceptor facilitates charge transfer by splitting the exciton.

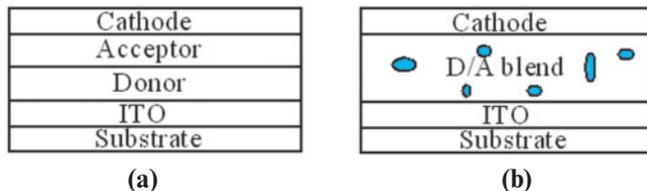


Fig. 25 (a) Bilayer and (b) bulk heterojunction solar cells.

Quantum Dot Solar Cells

As mentioned above, one way of increasing the conversion efficiency is to utilize tandem or cascaded solar cells that use two or more solar cells to increase the number of photons absorbed from the incident light.

Another approach to increasing the conversion efficiency is to utilize the hot carriers before they relax to the band edge via phonon emission.^{22,23} There are two fundamental ways of doing this: one is to extract hot carriers before they cool to enhance the photovoltage, the other is to utilize the energetic hot carriers to produce secondary (or more) EHPs through impact ionization to enhance the photocurrent.

The crucial point is to retard the relaxation of photogenerated carriers. Usually the energy of hot carriers is lost by multiphonon processes and heat is dissipated in the semiconductor. When the carriers in the semiconductor are confined by potential barriers in regions that are smaller than or comparable to their deBroglie wavelength or to the Bohr radius of excitons in the semiconductor bulk, i.e., in semiconductor quantum wells, quantum wires, and especially quantum dots (QDs), the relaxation of photogenerated carriers, especially hot carriers, may be markedly reduced by quantization effects in the semiconductor, and the rate of impact ionization can approach the rate of carrier cooling.

In order to achieve the former approach, the rates of photogenerated carrier separation, transport, and interfacial transfer across the contacts to the semiconductor must all be large compared to the rate of carrier cooling, as shown in Fig. 26a. In this configuration, the QDs are formed into an ordered 3-D array with inter-QD spacing sufficiently small so that strong electronic coupling occurs and minibands are formed to allow long-range electron transport. The QD array is placed in the intrinsic region of a $p^+ - i - n^+$ structure. The delocalized quantized 3-D miniband states could be expected to slow down the carrier cooling and permit the transport and collection of hot carriers at their respective p and n contacts to produce a higher photopotential in a solar cell.

The latter approach requires that the rate of impact ionization be greater than the rates of carrier cooling and other relaxation processes for hot carriers, as shown in Fig. 26b. Unlike bulk semiconductors, QDs possess a unique ability to generate multiple pairs of charge carriers with a single high-energy photon. In conventional bulk semiconductors, a single EHP is generated per absorbed photon. This means that both high- and low-energy photons create only a single pair of charge carriers (electron and hole). More simply, the extra energy of near-UV photons is not utilized fully when using bulk semiconductor films. In QDs, however, high-energy photons can produce multiple charge carriers by a process known as impact ionization, setting the stage for achieving photon-conversion efficiencies greater than 100%.

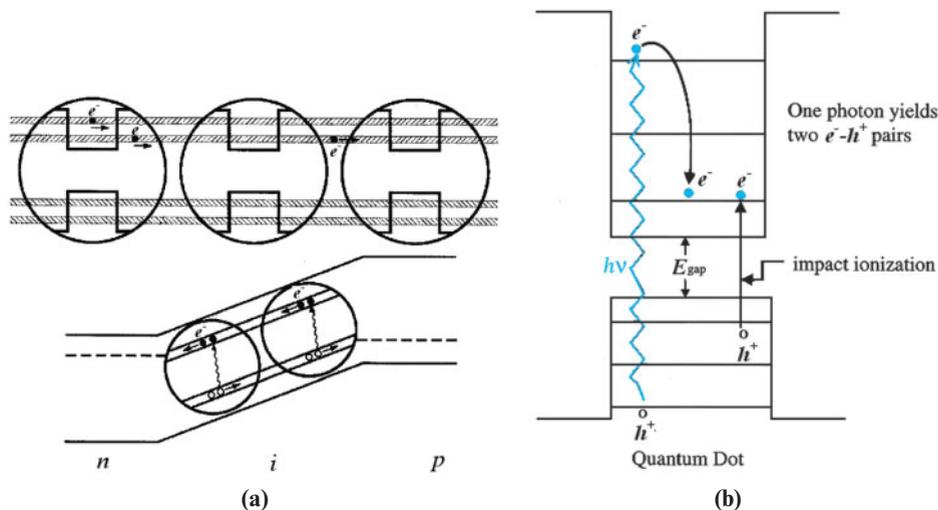


Fig. 26 (a) Hot carrier transport through the minibands of the QD array, resulting in a higher photopotential. (b) Enhanced efficiency could be achieved through impact ionization.^{22,23}

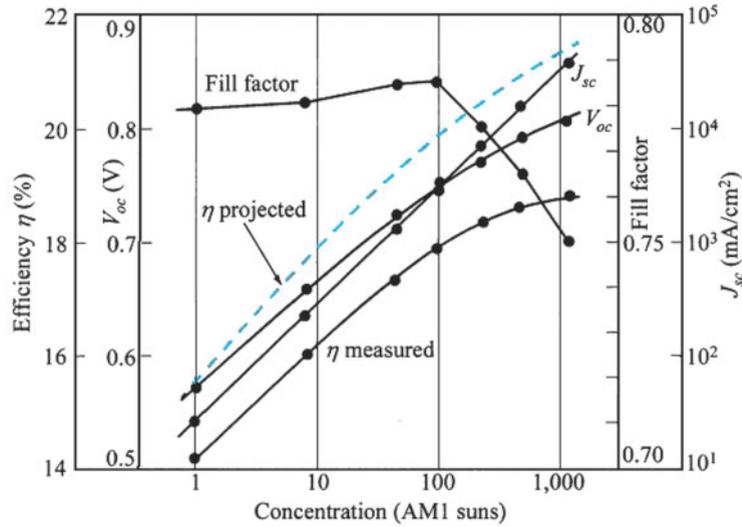


Fig. 27 Efficiency, open-circuit voltage, short-circuit current, and fill factor versus solar concentration.¹

However, hot-electron transport/collection and impact ionization cannot occur simultaneously; they are mutually exclusive and only one of these processes can be present in a given system. The QD solar cells not only promise high power-conversion efficiencies, but also offer spectral tunability, because the absorption properties of semiconductor quantum dots are size-dependent. QD solar cells have the potential to increase the maximum conversion efficiency up to 66%.

► 10.5 OPTICAL CONCENTRATION

Sunlight can be focused by using mirrors and lenses. Optical concentration offers an attractive and flexible approach to reducing high cell costs by substituting a concentrator area for much of the cell area. It also offers other advantages, such as a 20% increase in efficiency for a concentration of 1000 suns (an intensity of $963 \times 10^3 \text{ W/m}^2$). Figure 27 shows the measured results of a typical silicon solar cell mounted in a concentrated system.¹ Note that device performances improve as the concentration increases from one sun toward 1000 suns. The short-circuit current density increases linearly with concentration. The open-circuit voltage increases at a rate of 0.1 V per decade, while the fill factor varies slightly. The efficiency, which is the product of the foregoing three factors divided by the input power, increases at a rate of about 2% per decade. With a proper antireflection coating, we project an efficiency increase of 30% at 1000 suns. Therefore, one cell operated under 1000-sun concentration can produce the same power output as 1300 cells under one sun. Potentially, the optical concentration approach can replace expensive solar cells with less expensive concentrator materials and a related tracking and heat-removal system to minimize the overall system cost.

► SUMMARY

The operation of photodetectors and solar cells depends upon the absorption of photons. Photons are absorbed to create charge carriers. Photodetectors include photoconductors, photodiodes avalanche photodiodes phototransistor and etc.. They can convert optical signals into electrical signals. When photons are absorbed, EHPs are generated in the device that are subsequently separated by an electrical field to produce a photo-current flowing between the electrodes. Photodetectors are used for optical sensing and detection in opto-isolators and optical-fiber communication systems.

A solar cell is similar to a photodiode and has the same operational principle. However, the solar cell differs from a photodiode in that it is a large-area device and covers a wide range of the optical spectrum (solar radiation). Solar cells furnish the long-duration power supply for satellites. The solar cell is a major candidate for a terrestrial energy source because it can convert sunlight directly to electricity with good efficiency and is environmentally benign. Currently, the important solar cells are the highly efficient silicon PERL cell (24%), the GaInP/GaAs tandem cell (30%), and the low-cost thin-film microcrystalline a-Si solar cell (15%) and CIGS solar cell (19.8%). The next target for the solar cells is generally to provide higher efficiency and lower cost per watt of electricity generated. These so-called third-generation photovoltaic cells are under research and development.

► REFERENCES

1. S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, 3rd Ed., Wiley Interscience, Hoboken, 2007, Ch. 12-14.
2. S. R. Forrest, "Photodiodes for Long-Wavelength Communication Systems," *Laser Focus*, **18**, 81 (1982).
3. F. Capasso, W. T. Tsang, A. L. Hutchinson and G. F. Williams, "Enhancement of Electron Impact Ionisation in a Superlattice: A New Avalanche Photodiode with a Large Ionisation Rate Ratio," *Appl. Phys. Lett.*, **40**, 38 (1982).
4. D. M. Chapin, C. S. Fuller, and G. L. Pearson, "A New Silicon *p-n* Junction Photocell for Converting Solar Radiation into Electrical Power," *J. Appl. Phys.*, **25**, 676 (1954).
5. M. A. Green, "Solar Cells" in S. M. Sze, Ed., *Modern Semiconductor Device Physics*, Wiley Interscience, New York, 1998.
6. R. Hulstrom, R. Bird, and C. Riordan, "Spectral Solar Irradiance Data Sets for Selected Terrestrial Conditions," *Solar Cells*, **15**, 365 (1985).
7. C. H. Henry, "Limiting Efficiency of Ideal Single and Multiple Energy Gap Terrestrial Solar Cells," *J. Appl. Phys.*, **51**, 4494 (1980).
8. A. Luque, Ed., *Physical Limitation to Photovoltaic Energy Conversion*, IOP Press, Philadelphia, 1990.
9. M. A. Green, *Silicon Solar Cells: Advanced Principles and Practice*, Bridge Printery, Sydney, 1995.
10. M. Yamaguchi, T. Takamoto, and K. Araki, "Super high-efficiency multi-junction and concentrator solar cells," *Solar Energy Materials & Solar Cells*, **90**, 3068 (2006).
11. J. Macneil et al. "Recent Improvements in Very Large Area α -Si PV Module Manufacturing," in *Proc., 10th Euro. Photovolt. Sol. Energy Conf.*, Lisbon, 1188, 1991.
12. J. Yang, A. Banerjee, and S. Guha, "Triple-junction amorphous silicon alloy solar cell with 14.6% initial and 13.0% stable conversion efficiencies," *Appl. Phys. Lett.*, **70**, 2975 (1997).

13. A. V. Shah, J. Meier, E. Vallat-Sauvain, N. Wyrsh, U. Kroll, C. Droz, and U. Graf, "Material and solar cell research in microcrystalline silicon," *Solar Energy Materials & Solar Cells*, **78**, 469 (2003).
14. K. Sriprapa and P. Sichanugrist, "Amorphous/Microcrystalline Silicon Solar Cell Fabricated on Metal Substrate and Its Pilot Production," *Technical Digest of the International PVSEC-14*, Bangkok, Thailand 99, 2004.
15. K. Ramanathan, M. A. Contreras, C. L. Perkins, S. Asher, F. S. Hasoon, J. Keane, D. Young, M. Romero, W. Metzger, R. Noufi, J. Ward and A. Duda, "Properties of 19.2% Efficiency ZnO/CdS/CuInGaSe₂ Thin-film Solar Cells," *Prog. Photovolt: Res. Appl.*, **11**, 225 (2003).
16. I. Repins, M. A. Contreras, B. Egaas, C. DeHart, J. Scharf, C. L. Perkins, B. To, and R. Noufi, "19.9%-efficient ZnO/CdS/CuInGaSe₂ Solar Cell with 81.2% Fill Factor", *Prog. Photovolt: Res. Appl.*, **16**, 235 (2008).
17. A. M. Barnett and A. Rothwarf, "Thin-Film Solar Cells: A Unified Analysis of their Potential," *IEEE Trans. Electron Devices*, **ED-27**, 615 (1980).
18. M. A. Green, *Third Generation Photovoltaics Advanced Solar Energy Conversion*, Springer-Verlag, Berlin, 2003.
19. M. Grätzel, "Perspectives for Dye-sensitized Nanocrystalline Solar Cells", *Prog. Photovolt. Res. Appl.* **8**, 171 (2000).
20. M. Grätzel, "Photovoltaic performance and long-term stability of dye-sensitized meoscopic solar cells", *C. R. Chimie*, **9**, 578 (2006).
21. T. Y. Chu et al. "Highly efficient polycarbazole-based organic photovoltaic devices," *Appl. Phys. Lett.*, **95**, 063304 (2009).
22. A. J. Nozik, "Quantum Dot Solar Cells", *Physica E*, **14**, 115 (2002).
23. G. Conibeer, "Third-generation Photovoltaics", *Materials Today*, **10**, 42 (2007).

► **PROBLEMS (* DENOTES DIFFICULT PROBLEMS)**

FOR SECTION 10.1 PHOTODETECTORS

1. What is the ideal responsivity at a wavelength of 0.8 μm for (1) a GaAs homojunction, (2) an $\text{Al}_{0.34}\text{Ga}_{0.66}\text{As}$ homojunction, (3) a heterojunction formed between GaAs and $\text{Al}_{0.34}\text{Ga}_{0.66}\text{As}$, and (4) a two-terminal, monolithic, series-connected tandem photodetectors when the upper detector is made of $\text{Al}_{0.34}\text{Ga}_{0.66}\text{As}$ and the lower detector is made of GaAs?
2. A photoconductor with dimensions $L = 6 \text{ mm}$, $W = 2 \text{ mm}$, and $D = 1 \text{ mm}$ (Fig. 1) is placed under uniform radiation. The absorption of the light increases the current by 2.83 mA. A voltage of 10 V is applied across the device. As the radiation is suddenly cut off, the current falls, initially at a rate of 23.6 A/s. The electron and hole mobility are 3600 and 1700 $\text{cm}^2/\text{V}\cdot\text{s}$, respectively. Find (a) the equilibrium density of electron-hole pairs generated under radiation, (b) the minority-carrier lifetime, and (c) the excess density of electrons and holes remaining 1 ms after the radiation is cut off.

3. Calculate the gain and current generated when $1 \mu\text{W}$ of optical power with $h\nu = 3 \text{ eV}$ shines onto a photoconductor of $\eta = 0.85$ and a minority carrier lifetime of 0.6 ns . The material has an electron mobility of $3000 \text{ cm}^2/\text{V}\cdot\text{s}$, the electric field is 5000 V/cm , and $L = 10 \mu\text{m}$.
4. The absorption coefficient is $4 \times 10^4 \text{ cm}^{-1}$ and a surface reflectivity is 0.1 for a Si p - n photodiode. The thicknesses of p and depletion regions are both $1 \mu\text{m}$ and the internal quantum efficiency is 0.8 . Calculate the external quantum efficiency.
- *5. Show that the quantum efficiency η of a p - i - n photodetector is related to the responsivity $\mathcal{R} = (I_p / P_{opt})$ at a wavelength λ (μm) by the equation $\mathcal{R} = \eta\lambda/1.24$.
6. Calculate the responsivity and hence the photocurrent when $5 \mu\text{W}$ of optical power at a wavelength of $1.1 \mu\text{m}$ is incident on the Si p - n photodiode of Prob. 4.
- *7. A silicon n^+ - p - π - p^+ avalanche photodiode operated at $0.8 \mu\text{m}$ has a p -layer of $3 \mu\text{m}$ and a π -layer $9 \mu\text{m}$ thick. The biasing voltage must be high enough to cause avalanche breakdown in the p -region and velocity saturation in the π -region. Find the minimum required biasing voltage and the corresponding doping concentration of the p -region. Estimate the transit time of the device.
8. The width of i layer in a Si p - i - n photodiode shown in Fig. 4 is $20 \mu\text{m}$. The p^+ -layer is $0.1 \mu\text{m}$. The p - i - n photodiode is operated under reverse bias of 100 V and illuminated with a very short optical pulse of wavelength 900 nm . What is the transit time of the photocurrent if absorption occurs over the whole i layer?
9. For a photodiode, we need a sufficiently wide depletion layer to absorb most the incoming light, but not too wide to limit the frequency response. Find the optimum depletion-layer thickness for Si photodiode having a modulation frequency of 10 GHz .

FOR SECTION 10.2 SOLAR CELLS

10. The sun (radius $r_s = 695,990 \text{ km}$) can be modeled as an ideal black-body at 6000 K . Calculate the Earth's surface temperature, assuming its temperature is uniform over its entire surface and that the sun is the only source providing energy to it (the mean sun-Earth distance d_{es} is $149,597,871 \text{ km}$). Assume the following models for the Earth's radiative properties: The rate at which a blackbody emits or absorbs energy is $P = \sigma AT^4$, A is the surface area, T is the temperature of that area, and σ is the Stefan-Boltzmann constant.
 - (a) Calculate the ideal black-body properties for the Earth.
 - (b) The absorptance averaged over the sun's energy spectrum is 0.7 , and the Earth's averaged radiative emissivity is 0.6 due to greenhouse gases.
 - (c) How much change in the latter absorptance is required to increase the Earth's temperature by $2 \text{ }^\circ\text{C}$?

11. A p - n junction photodiode can be operated under photovoltaic conditions similar to those for a solar cell. The current-voltage characteristics of a photodiode under illumination are also similar (Fig. 14). State three major differences between a photodiode and a solar cell.
12. A Si solar cell has a short-circuit current of 90 mA and an open-circuit voltage of 0.75 V under solar illumination. The fill factor is 0.8. What is the maximum power delivered to a load by this cell?
13. A solar cell with an area of 4 cm^2 has the I-V characteristics under the illumination of 600 Wcm^{-2} shown in Fig. 15. It drives a load of 5Ω . Calculate the power delivered to the load and the efficiency of the solar cell in this circuit.
- *14 Consider a silicon p - n junction solar cell of area 2 cm^2 . If the dopings of the solar cell are $N_A = 1.7 \times 10^{16} \text{ cm}^{-3}$ and $N_D = 5 \times 10^{19} \text{ cm}^{-3}$, and given $\tau_n = 10 \mu\text{s}$, $\tau_p = 0.5 \mu\text{s}$, $D_n = 9.3 \text{ cm}^2/\text{s}$, $D_p = 2.5 \text{ cm}^2/\text{s}$, and $I_L = 95 \text{ mA}$, (a) calculate and plot the I-V characteristics of the solar cell, (b) calculate the open-circuit voltage, and (c) determine the maximum output power of the solar cell, all at room temperature.
- *15. At 300 K, an ideal solar cell has a short-circuit current of 3 A and an open-circuit voltage of 0.6 V. Calculate and sketch its power output as a function of operating voltage and find its fill factor from this power output.
16. For the solar cell shown in Fig. 17, find the relative maximum power output for a R_L of 0 and 5Ω .
17. The absorption coefficients of amorphous Si and CIGS are approximately 10^4 cm^{-1} and 10^5 cm^{-1} at $h\nu = 1.7 \text{ eV}$, respectively. Determine the amorphous Si and CIGS thicknesses for each solar cell so that 90 % of the photons are absorbed
18. For solar cells operated under solar-concentration conditions (Fig. 27 with the measured η), how many such solar cells operated under one-sun conditions are needed to produce the same power output as one cell operated under a 10-sun, 100-sun, or 1000- sun concentration?

Crystal Growth and Epitaxy

- ▶ 11.1 SILICON CRYSTAL GROWTH FROM THE MELT
 - ▶ 11.2 SILICON FLOAT-ZONE PROCESS
 - ▶ 11.3 GaAs CRYSTAL-GROWTH TECHNIQUES
 - ▶ 11.4 MATERIAL CHARACTERIZATION
 - ▶ 11.5 EPITAXIAL-GROWTH TECHNIQUES
 - ▶ 11.6 STRUCTURES AND DEFECTS IN EPITAXIAL LAYERS
 - ▶ SUMMARY
-

As discussed in Chapter 1, the two most important semiconductors for discrete devices and integrated circuits are silicon and gallium arsenide. In this chapter we describe the common techniques for growing single crystals of these two semiconductors. The starting materials, silicon dioxide for a silicon wafer and gallium and arsenic for a gallium arsenide wafer, are chemically processed to form a high-purity polycrystalline semiconductor from which single crystals are grown. The single-crystal ingots are shaped to define the diameter of the material and sawed into wafers. These wafers are etched and polished to provide smooth, specular surfaces on which devices will be made.

A technology closely related to crystal growth involves the growth of single-crystal semiconductor layers on a single-crystal semiconductor substrate. This is called *epitaxy*, from the Greek words epi (meaning “on”) and taxis (meaning “arrangement”). The epitaxial layer and the substrate materials may be the same, giving rise to *homoepitaxy*. For example, an n -type silicon can be grown epitaxially on an n^+ -silicon substrate. On the other hand, if the epitaxial layer and the substrate are chemically and often crystallographically different, we have *heteroepitaxy*, such as the epitaxial growth of $\text{Al}_x\text{Ga}_{1-x}\text{As}$ on GaAs.

Specifically, we cover the following topics:

- Basic techniques to grow silicon and GaAs single-crystal ingots.
- Wafer-shaping steps from ingots to polished wafers.
- Wafer characterization in term of its electrical and mechanical properties.
- Basic techniques of epitaxy, that is, growing a single-crystal layer on a single-crystal substrate.
- Structures and defects of lattice-matched and strained-layer epitaxial growth.

▶ 11.1 SILICON CRYSTAL GROWTH FROM THE MELT

The basic technique for silicon crystal growth from the melt, which is material in liquid form, is the Czochralski technique.^{1,2} The Czochralski process is the most common and most advanced method for semiconductor single crystal growth. The process is named after Polish scientist Jan Czochralski, who discovered the method in 1916

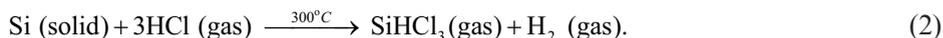
while investigating the crystallization rates of metals. A substantial percentage (> 90%) of the silicon crystals for the semiconductor industry are prepared by the Czochralski technique and virtually all the silicon used for fabricating integrated circuits is prepared by this technique.

11.1.1 Starting Material

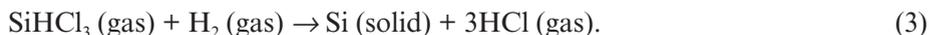
The starting material for silicon is a relatively pure form of sand (SiO_2) called quartzite. This is placed in a furnace with various forms of carbon (coal, coke, and wood chips). Although a number of reactions take place in the furnace, the overall reaction is



This process produces metallurgical-grade silicon with a purity of about 98%. Next, the silicon is pulverized and treated with hydrogen chloride (HCl) to form trichlorosilane (SiHCl_3):



The trichlorosilane is a liquid at room temperature (boiling point 32°C). Fractional distillation of the liquid removes the unwanted impurities. The purified SiHCl_3 is then used in a hydrogen reduction reaction to prepare the electronic-grade silicon (EGS):



This reaction takes place in a reactor containing a resistance-heated silicon rod, that serves as the nucleation point for the deposition of silicon. The EGS, a polycrystalline material of high purity, is the raw material used to prepare device-quality, single-crystal silicon. Pure EGS generally has impurity concentrations in the parts-per-billion range.

11.1.2 The Czochralski Technique

The Czochralski technique uses an apparatus called a crystal puller shown in Fig. 1a. The puller has three main components: (a) a furnace, which includes a fused-silicon (SiO_2) crucible, a graphite susceptor, a rotation mechanism (clockwise as shown), a heating element, and a power supply. The crucible rotates during the growth to prevent the formation of local hot or cold regions; (b) a crystal-pulling mechanism that includes a seed holder and a rotation mechanism (counter-clockwise); and (c) an ambient control that includes a gas source (such as argon to prevent contamination of the molten silicon), a flow control, and an exhaust system. In addition, the puller has an overall microprocessor-based control system to control process parameters such as temperature, crystal diameter, pull rate, and rotation speeds, as well as to permit programmed process steps. Also, various sensors and feedback loops allow the control system to respond automatically, reducing operator intervention.

In the crystal-growing process, polycrystalline silicon (EGS) is placed in the crucible shown in Fig. 1b and the furnace is heated above the melting temperature of silicon (1412°C). A suitably oriented seed crystal (e.g., $\langle 111 \rangle$) is suspended over the crucible in a seed holder. The seed is inserted into the melt. Part of it melts, but the tip of the remaining seed crystal still touches the liquid surface. It is then slowly withdrawn from the melt, as shown in Fig. 1c. The molten silicon adhering to the crystal freezes or solidifies, using the crystal of the seed crystal as a template. Progressive freezing at the solid-liquid interface yields a large single crystal. The desired impurity concentration is obtained by adding impurities to the melt in the form of heavily doped silicon prior to crystal growth. Figure 2 shows silicon ingot weight and wafer diameter will reach 450 Kg and 450 mm (18 in.) around 2015. The relentless increase is mainly to reduce the processing cost per unit area

11.1.3 Distribution of Dopant

In crystal growth, a known amount of dopant is added to the melt to obtain the desired doping concentration in the grown crystal. For silicon, boron and phosphorus are the most common dopants for *p*- and *n*-type materials, respectively.

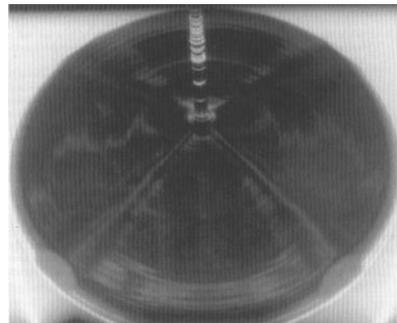
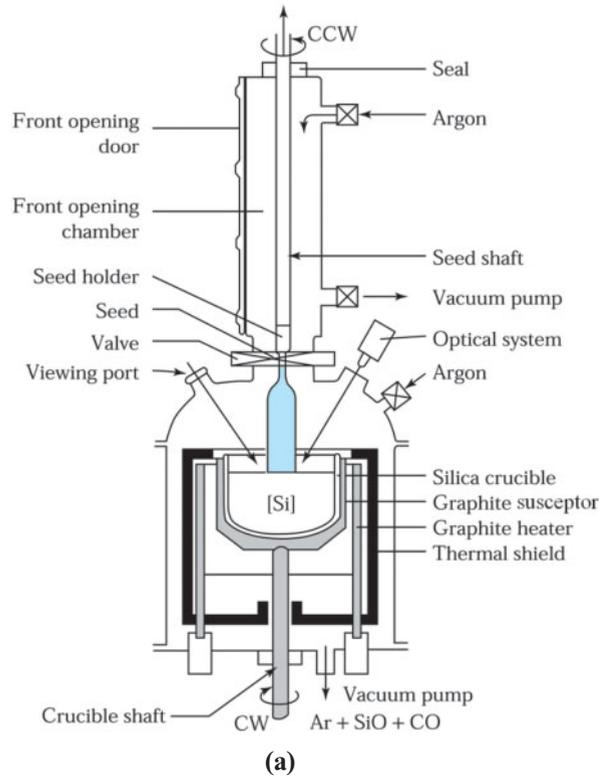


Fig. 1 (a) Schematic drawing of the Czochralski crystal puller. CW, clockwise; CCW, counter clockwise. (b) Photograph of polystalline silicon in a silica crucible. (c) Photograph of a 200 mm diameter, (100)-oriented Si crystal being pulled from the melt. (Photographs courtesy of Taisil Electronic Materials Corp., Taiwan.)

As a crystal is pulled from the melt, the doping concentration incorporated into the crystal (solid) is usually different from the doping concentration of the melt (liquid) at the interface. The ratio of these two concentrations is defined as the *equilibrium segregation coefficient* k_0 :

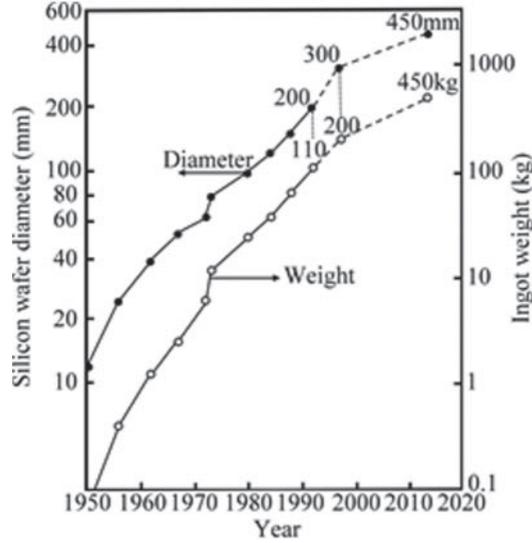


Fig. 2 Increase of wafer diameter and weight of Czochralski-grown silicon ingot from 1950 to 2000 and projected to 2015.

$$k_0 = \frac{C_s}{C_l}, \quad (4)$$

where C_s and C_l are, respectively, the equilibrium concentrations of the dopant in the solid and liquid near the interface. Table 1 lists values of k_0 for the commonly used dopants for silicon. Note that most values are below 1, which means that during growth the dopants are rejected into the melt. Consequently, the melt becomes progressively enriched with the dopant as the crystal grows.

Consider a crystal being grown from a melt having an initial weight M_0 with an initial doping concentration C_0 in the melt (i.e., the weight of the dopant per 1 g of melt). At a given point of growth when a crystal of weight M has been grown, the amount of dopant remaining in the melt (by weight) is S . For an incremental amount of the crystal with weight dM , the corresponding reduction in the dopant ($-dS$) from the melt is $C_s dM$, where C_s is the doping concentration in the crystal (by weight):

$$-dS = C_s dM. \quad (5)$$

Now, the remaining weight of the melt is $M_0 - M$, and the doping concentration in the liquid (by weight), C_l , is given by

$$C_l = \frac{S}{M_0 - M}. \quad (6)$$

Combining Eqs. 5 and 6 and substituting $C_s/C_l = k_0$ yields

$$\frac{dS}{S} = -k_0 \left(\frac{dM}{M_0 - M} \right). \quad (7)$$

Given the initial weight of the dopant, $C_0 M_0$, we can integrate Eq. 7:

$$\int_{C_0 M_0}^S \frac{dS}{S} = k_0 \int_0^M \frac{-dM}{M_0 - M}. \quad (8)$$

TABLE 1 EQUILIBRIUM SEGREGATION COEFFICIENTS FOR DOPANTS IN SI

| Dopant | k_0 | Type | Dopant | k_0 | Type |
|--------|--------------------|----------|--------|----------------------|----------------|
| B | 8×10^{-1} | <i>p</i> | As | 3.0×10^{-1} | <i>n</i> |
| Al | 2×10^{-3} | <i>p</i> | Sb | 2.3×10^{-2} | <i>n</i> |
| Ga | 8×10^{-3} | <i>p</i> | Te | 2.0×10^{-4} | <i>n</i> |
| In | 4×10^{-4} | <i>p</i> | Li | 1.0×10^{-2} | <i>n</i> |
| O | 1.25 | <i>n</i> | Cu | 4.0×10^{-4} | — ^a |
| C | 7×10^{-2} | <i>n</i> | Au | 2.5×10^{-5} | — ^a |
| P | 0.35 | <i>n</i> | | | |

^aDeep-lying impurity level.

Solving Eq. 8 and combining with Eq. 6 gives

$$C_s = k_0 C_0 \left(1 - \frac{M}{M_0} \right)^{k_0 - 1} \quad (9)$$

Figure 3 illustrates the doping distribution as a function of the fraction solidified (M/M_0) for several segregation coefficients.^{3,4} As crystal growth progresses, a composition initially at $k_0 C_0$ will increase continually for $k_0 < 1$ and decrease continually for $k_0 > 1$. When $k_0 \cong 1$, a uniform impurity distribution can be obtained.

► EXAMPLE 1

A silicon ingot that should contain 10^{16} boron atoms/cm³ is to be grown by the Czochralski technique. What concentration of boron atoms should be in the melt to give the required concentration in the ingot? If the initial load of silicon in the crucible is 60 kg, how many grams of boron (atomic weight 10.8) should be added? The density of molten silicon is 2.53 g/cm³.

SOLUTION Table 1 shows that the segregation coefficient k_0 for boron is 0.8. We assume that $C_s = k_0 C_l$ throughout the growth. Thus, the initial concentration of boron in the melt should be

$$\frac{10^{16}}{0.8} = 1.25 \times 10^{16} \text{ boron atoms/cm}^3.$$

Since the amount of boron concentration is so small, the volume of melt can be calculated from the weight of silicon. Therefore, the volume of 60 kg of silicon is

$$\frac{60 \times 10^3}{2.53} = 2.37 \times 10^4 \text{ cm}^3.$$

The total number of boron atoms in the melt is

$$1.25 \times 10^{16} \text{ atoms/cm}^3 \times 2.37 \times 10^4 \text{ cm}^3 = 2.96 \times 10^{20} \text{ boron atoms,}$$

so that

$$\begin{aligned} \frac{2.96 \times 10^{20} \text{ atoms} \times 10.8 \text{ g/mol}}{6.02 \times 10^{23} \text{ atoms/mol}} &= 5.31 \times 10^{-3} \text{ g of boron,} \\ &= 5.31 \text{ mg of boron.} \end{aligned}$$

Note the small amount of boron needed to dope such a large load of silicon. ◀

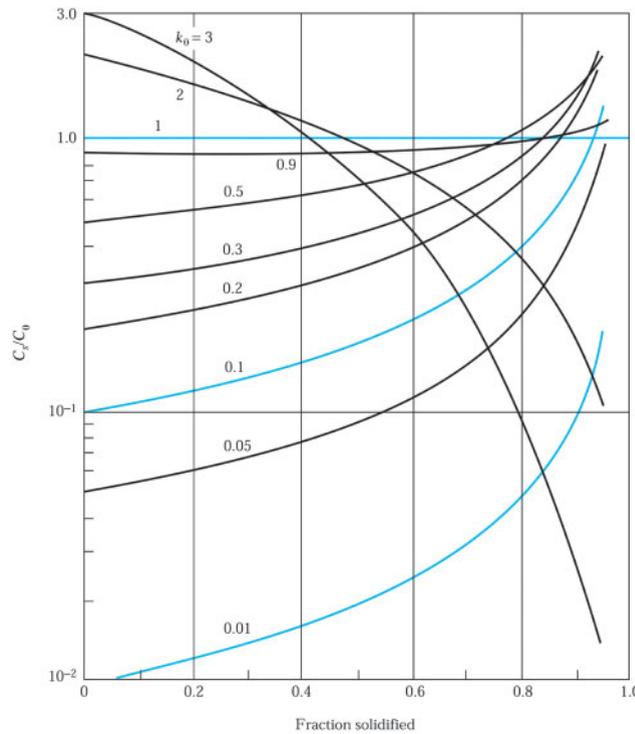


Fig. 3 Curves for growth from the melt showing the doping concentration in a solid as a function of the fraction solidified.⁴

11.1.4 Effective Segregation Coefficient

While the crystal is growing, dopants are constantly being rejected into the melt (for $k_0 < 1$). If the rejection rate is higher than the rate at which the dopant can be transported away by diffusion or stirring, then a concentration gradient will be developed at the interface, as illustrated in Fig. 4. The segregation coefficient (given in Section 11.1.3) is $k_0 = C_s/C_l(0)$. We can define an effective segregation coefficient k_e that is the ratio of C_s and the impurity concentration far away from the interface:

$$k_e \equiv \frac{C_s}{C_l} \quad (10)$$

Consider a small, virtually stagnant layer of melt with width δ in which the only flow is that required to replace the crystal being withdrawn from the melt. Outside this stagnant layer, the doping concentration has a constant value C_l . Inside the layer, the doping concentration can be described by the continuity equation (Eq. 59) derived in Chapter 2. At steady state, the only significant terms are the second and third terms on the right-hand side (we replace n_p by C and $\mu_n \mathcal{E}$ by v):

$$0 = v \frac{dC}{dx} + D \frac{d^2C}{dx^2}, \quad (11)$$

where D is the dopant diffusion coefficient in the melt, v is the crystal growth velocity, and C is the doping concentration in the melt.

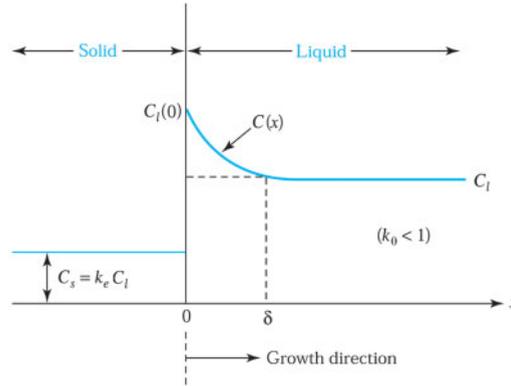


Fig. 4 Doping distribution near the solid-melt interface.

The solution of Eq. 11 is

$$C = A_1 e^{-vx/D} + A_2 \quad (12)$$

where A_1 and A_2 are constants to be determined by the boundary conditions. The first boundary condition is that $C = C_l(0)$ at $x = 0$. The second boundary condition is the conservation of the total amount of dopants; that is, the sum of the dopant fluxes at the interface must be zero. By considering the diffusion of dopant atoms in the melt (neglecting diffusion in the solid), we have

$$D \left(\frac{dC}{dx} \right)_{x=0} + [C_l(0) - C_s]v = 0. \quad (13)$$

Substituting these boundary conditions into Eq. 12 and noting that $C = C_l$ at $x = \delta$ gives

$$e^{-v\delta/D} = \frac{C_l - C_s}{C_l(0) - C_s}. \quad (14)$$

Therefore,

$$k_e \equiv \frac{C_s}{C_l} = \frac{k_0}{k_0 + (1 - k_0)e^{-v\delta/D}}. \quad (15)$$

The doping distribution in the crystal is given by the same expression as in Eq. 9, except that k_0 is replaced by k_e . Values of k_e are larger than those of k_0 and can approach 1 for large values of the growth parameter $v\delta/D$. Uniform doping distribution ($k_e \rightarrow 1$) in the crystal can be obtained by employing a high pull rate and low rotation speed (since δ is inversely proportional to the rotation speed). Another approach to achieve uniform doping is to add ultrapure polycrystalline silicon continuously to the melt so that the initial doping concentration is maintained.

► 11.2 SILICON FLOAT-ZONE PROCESS

The float-zone process can be used to grow silicon that has lower contamination than that normally obtained from the Czochralski technique. A schematic setup of the float-zone process is shown in Fig. 5a. A high-purity polycrystalline rod with a seed crystal at the bottom is held in a vertical position and rotated. The rod is enclosed in a quartz envelope within which an inert atmosphere (argon) is maintained. During the operation, a small zone (a few centimeters in length) of the crystal is kept molten by a radio-frequency heater, which is moved from the seed upward so that this *floating zone* traverses the length of the rod. The molten silicon is retained by surface

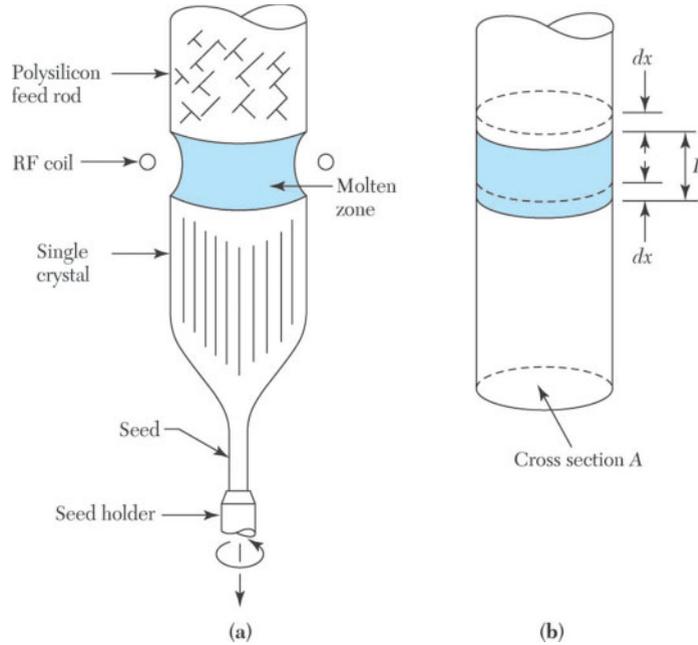


Fig. 5 Float-zone process. (a) Schematic setup. (b) Simple model for doping evaluation, end of the rod. From Eq. 16a we obtain

tension between the melting and growing solid-silicon faces. As the floating zone moves upward, a single-crystal silicon freezes at the zone's retreating end and grows as an extension of the seed crystal. Materials with higher resistivities can be obtained from the float-zone process than from the Czochralski process because it can be used to purify the crystal more easily. Furthermore, since no crucible is used in the float-zone process, there is no contamination from the crucible (as with Czochralski growth). At the present time, float-zone crystals are used mainly for high-power, high-voltage devices, where high-resistivity materials are required.

To evaluate the doping distribution of a float-zone process, consider the simplified model shown in Fig. 5b. The initial, uniform doping concentration in the rod is C_0 (by weight). L is the length of the molten zone at a distance x along the rod, A the cross-sectional area of the rod, ρ_d the specific density of silicon, and S the amount of dopant present in the molten zone. As the zone traverses a distance dx , the amount of dopant added to it at its advancing end is $C_0\rho_d A dx$, whereas the amount of dopant removed from it at the retreating end is $k_e(S dx/L)$, where k_e is the effective segregation coefficient. Thus,

$$dS = C_0\rho_d A dx - \frac{k_e S}{L} dx = \left(C_0\rho_d A - \frac{k_e S}{L} \right) dx, \quad (16)$$

so that

$$\int_0^x dx = \int_{S_0}^S \frac{dS}{C_0\rho_d A - (k_e S/L)}, \quad (16a)$$

where $S_0 = C_0\rho_d AL$ is the amount of dopant in the zone when it was first formed at the front end of the rod. From Eq 16a, we obtain

$$\exp\left(\frac{k_e x}{L}\right) = \frac{C_0\rho_d A - (k_e S_0/L)}{C_0\rho_d A - (k_e S/L)} \quad (17)$$

or

$$S = \frac{C_0 A \rho_d L}{k_e} [1 - (1 - k_e) e^{-k_e x/L}]. \quad (17a)$$

Since C_s (the doping concentration in the crystal at the retreating end) is given by $C_s = k_e(S/A\rho_dL)$, we have

$$C_s = C_0 [1 - (1 - k_e) e^{-k_e x/L}]. \quad (18)$$

Figure 6 shows the doping concentration versus the solidified zone length for various values of k_e .

These two crystal growth techniques can also be used to remove impurities. Comparison of Fig. 6 with Fig. 3 shows that a single pass in the float-zone process does not produce as much purification as a single Czochralski growth. For example, for $k_0 = k_e = 0.1$, C_s/C_0 is smaller over most of the solidified ingot made by the Czochralski growth. However, multiple float-zone passes can be performed on a rod much more easily than a crystal can be grown, the end region cropped off, and regrown from the melt. Figure 7 shows the impurity distribution for an element with $k_e = 0.1$ after a number of successive passes of the zone along the length of the rod.⁴ Note that there is a substantial reduction of impurity concentration in the rod after each pass. Therefore, the float-zone process is ideally suited for crystal purification. This process is also called the zone-refining technique, and it can provide a very high purity level of the raw material.

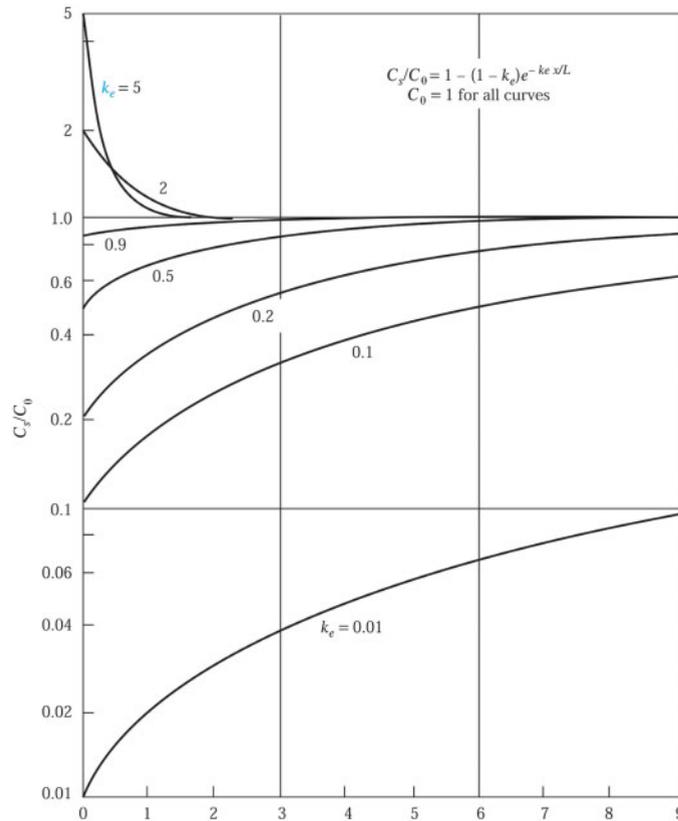


Fig. 6 Curves for the float-zone process showing doping concentration in the solid as a function of solidified zone lengths.⁴

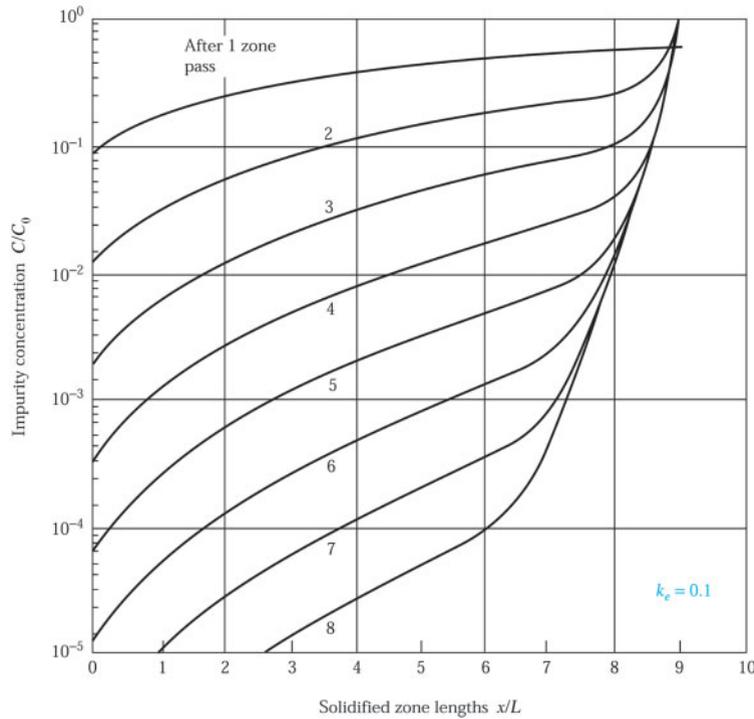


Fig. 7 Relative impurity concentration versus zone length for a number of passes. L denotes the zone length.⁴

If it is desirable to dope the rod rather than purify it, consider the case in which all the dopants are introduced in the first zone ($S_0 = C_1 A \rho_d L$) and the initial concentration C_0 is negligibly small. Equation 17 gives

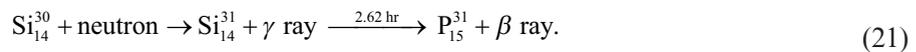
$$S_0 = S \exp\left(\frac{k_e x}{L}\right). \quad (19)$$

Since $C_s = k_e (S/A \rho_d L)$, we obtained from Eq. 19

$$C_s = k_e C_1 e^{-k_e x/L}. \quad (20)$$

Therefore, if $k_e x/L$ is small, C_s will remain nearly constant with distance except at the end that is last to solidify.

For certain switching devices, such as high-voltage thyristors discussed in Chapter 4, large chip areas are used, frequently an entire wafer for a single device. This size imposes stringent requirements on the uniformity of the starting material. To obtain homogeneous distribution of dopants, we use a float-zone silicon slice that has an average doping concentration well below the required amount. The slice is then irradiated with thermal neutrons. This process, called *neutron irradiation*, gives rise to fractional transmutation of silicon into phosphorus and dopes the silicon *n*-type:



The half-life of the intermediate element Si_{14}^{31} is 2.62 hours. Because the penetration depth of neutrons in silicon is about 100 cm, doping is very uniform throughout the slice. Figure 8 compares the lateral resistivity distributions in conventionally doped silicon and in silicon doped by neutron irradiation.⁵ Note that the resistivity variations for the neutron-irradiated silicon are much smaller than for the conventionally doped silicon.

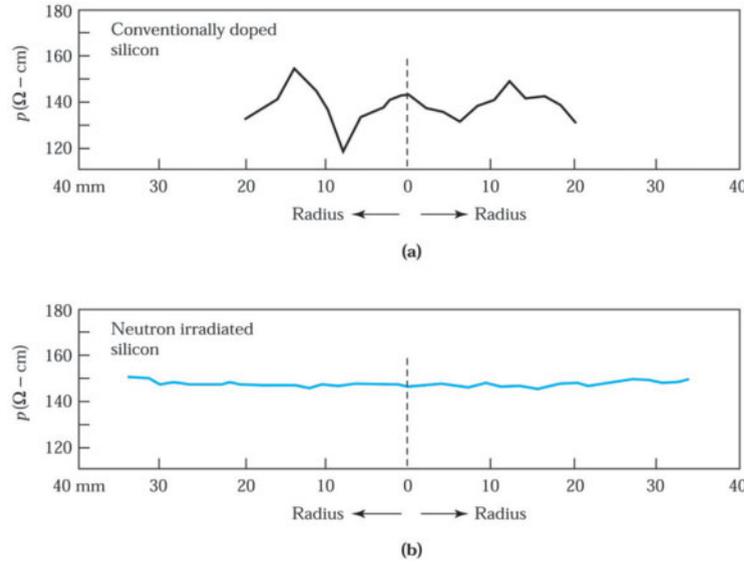


Fig. 8 (a) Typical lateral resistivity distribution in conventionally doped silicon. (b) Silicon doped by neutron irradiation.⁵

► 11.3 GaAs CRYSTAL-GROWTH TECHNIQUES

11.3.1 Starting Materials

The starting materials for the synthesis of polycrystalline gallium arsenide are elemental, chemically pure gallium and arsenic. Because gallium arsenide is a combination of two materials, its behavior is different from that of a single material such as silicon. The behavior of a combination can be described by a *phase diagram*. A phase is a state (e.g., solid, liquid, or gaseous) in which a material may exist. A phase diagram shows the relationship between the two components, gallium and arsenic, as a function of temperature.

Figure 9 shows the phase diagram of the gallium-arsenic system. The abscissa represents various compositions of the two components in terms of atomic percent (lower scale) or weight percent (upper scale).^{6,7} Consider a melt that is initially of composition x (e.g., 85 atomic percent arsenic shown in Fig. 9). When the temperature is lowered, its composition will remain fixed until the *liquidus* line is reached. At the point (T_l, x) , material of 50 atomic percent arsenic (i.e., gallium arsenide) will begin to solidify.

► EXAMPLE 2

In Fig. 9, consider a melt of initial composition C_m (weight percent scale) that is cooled from T_a (on the liquidus line) to T_b . Find the fraction of the melt that will be solidified.

SOLUTION At T_b , M_l is the weight of the liquid, M_s the weight of the solid (i.e., GaAs), and C_l and C_s are the concentrations of dopant in the liquid and the solid, respectively. Therefore, the weights of arsenic in the liquid and solid are $M_l C_l$ and $M_s C_s$, respectively. Because the total arsenic weight is $(M_l + M_s)C_m$, we have

$$M_l C_l + M_s C_s = (M_l + M_s) C_m$$

or

$$\frac{M_s}{M_l} = \frac{\text{weight of GaAs at } T_b}{\text{weight of liquid at } T_b} = \frac{C_m - C_l}{C_s - C_m} = \frac{s}{l},$$

where s and l are the lengths of the two lines measured from C_m to the liquidus and solidus line, respectively. As can be seen from Fig. 9, about 10% of the melt is solidified. ◀

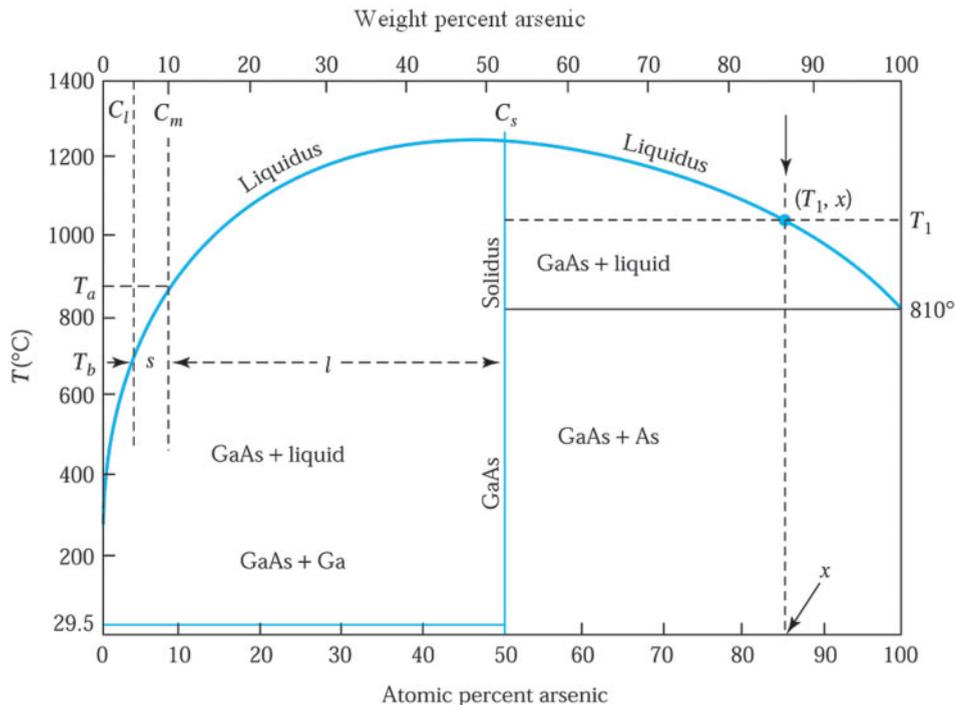


Fig. 9 Phase diagram for the gallium-arsenic system.⁶

Unlike silicon, which has a relatively low vapor pressure at its melting point ($\sim 10^{-6}$ atm at 1412°C), arsenic has much higher vapor pressures at the melting point of gallium arsenide (1240°C). In its vapor phase, arsenic has As_2 and As_4 as its major species. Figure 10 shows the vapor pressures of gallium and arsenic along the liquidus curve.⁸ Also shown for comparison is the vapor pressure of silicon. The vapor pressure curves for gallium arsenide are double valued. The dashed curves are for arsenic-rich gallium arsenide melt (right side of liquidus line in Fig. 9), and the solid curves are for gallium-rich gallium arsenide melt (left side of liquidus line in Fig. 9). Because there is more arsenic in an arsenic-rich melt than in a gallium-rich melt, more arsenic (As_2 and As_4) will be vaporized from the arsenic-rich melt, thus resulting in a higher vapor pressure. A similar argument can explain the higher vapor pressure of gallium in a gallium-rich melt. Note that long before the melting point is reached, the surface layers of liquid gallium arsenide may decompose into gallium and arsenic. Since the vapor pressures of gallium and arsenic are different, there is a preferential loss of the more volatile arsenic species, and the liquid becomes gallium rich.

To synthesize gallium arsenide, an evacuated, sealed quartz-tube system with a two-temperature furnace is commonly used. The high-purity arsenic is placed in a graphite boat and heated to $610^\circ\text{--}620^\circ\text{C}$, whereas the high-purity gallium is placed in another graphite boat and heated to slightly above the gallium arsenide melting temperature ($1240^\circ\text{--}1260^\circ\text{C}$). Under these conditions, an overpressure of arsenic is established (a) to cause the transport of arsenic vapor to the gallium melt, converting it into gallium arsenide, and (b) to prevent decomposition of the gallium arsenide while it is being formed in the furnace. When the melt cools, a high-purity polycrystalline gallium arsenide results. This serves as the raw material to grow single-crystal gallium arsenide.⁷

11.3.2 Crystal-Growth Techniques

There are two techniques for GaAs crystal growth: the Czochralski technique and the Bridgman technique. Most gallium arsenide is grown by the Bridgman technique. However, the Czochralski technique is more popular for the growth of larger-diameter GaAs ingots.

For Czochralski growth of gallium arsenide, the basic puller is identical to that for silicon. However, to prevent decomposition of the melt during crystal growth, a liquid encapsulation method is employed. The liquid

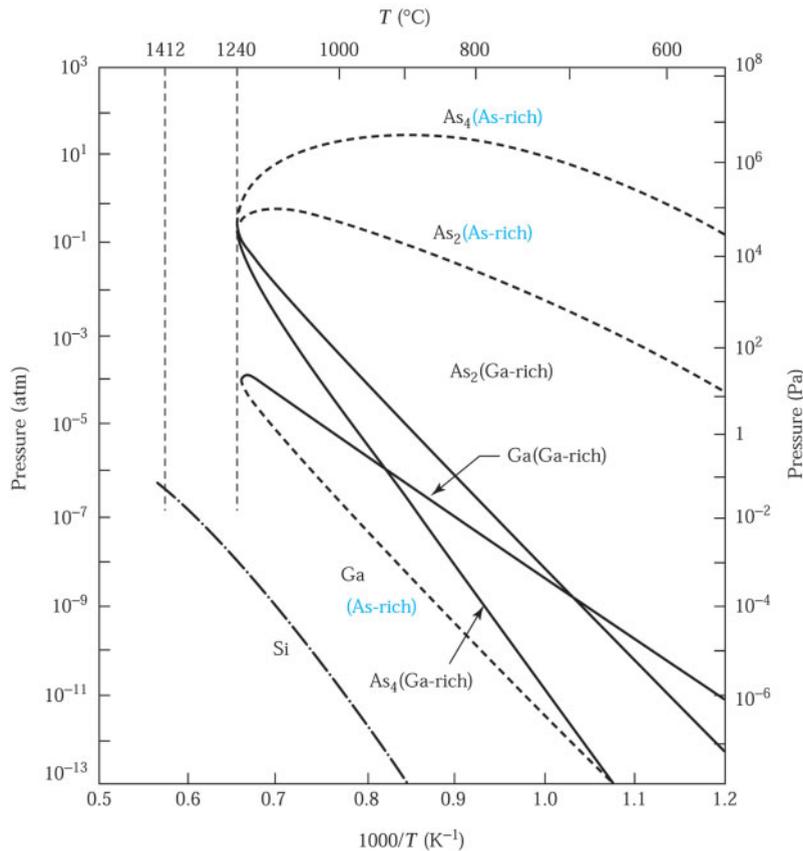


Fig. 10 Partial pressure of gallium and arsenic over gallium arsenide as a function of temperature.⁸ Also shown is the partial pressure of silicon.

encapsulant is a molten boron trioxide (B_2O_3) layer about 1 cm thick. Molten boron trioxide is inert to the gallium arsenide surface and serves as a cap to cover the melt. This cap prevents decomposition of the gallium arsenide as long as the pressure on its surface is higher than 1 atm (760 Torr). Because boron trioxide dissolves silicon dioxide, the fused-silica crucible is replaced with a graphite crucible.

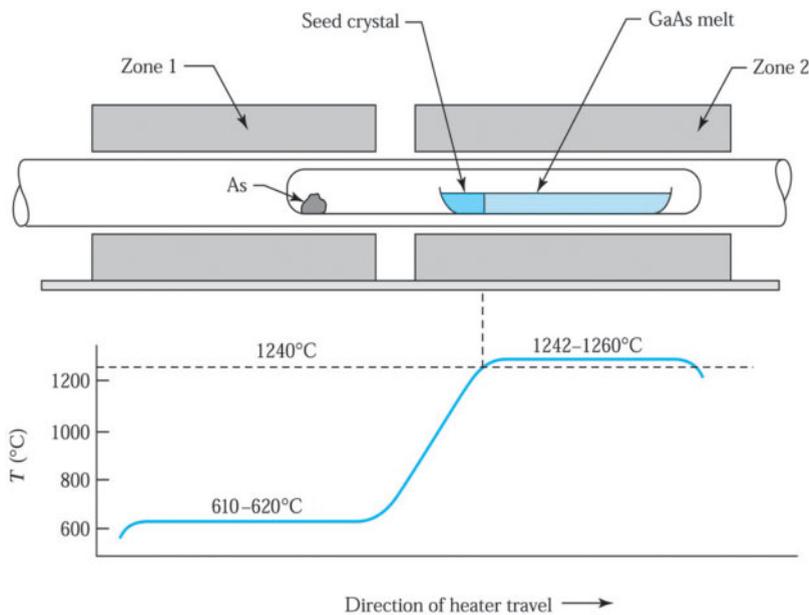
To obtain the desired doping concentration in the grown crystal of GaAs, cadmium and zinc are commonly used for *p*-type materials, whereas selenium, silicon, and tellurium are used for *n*-type materials. For semiinsulating GaAs, the material is undoped. The equilibrium segregation coefficients for dopants in GaAs are listed in Table 2. As in Si, most of the segregation coefficients are less than 1. The expressions derived previously for Si are equally applicable to GaAs (Eqs. 4 to 15).

Figure 11 shows a Bridgman system in which a two-zone furnace is used for growing single-crystal gallium arsenide. The left-hand zone is held at a temperature ($\sim 610^\circ\text{C}$) to maintain the required overpressure of arsenic, whereas the right-hand zone is held just above the melting point of gallium arsenide (1240°C). The sealed tube is made of quartz and the boat is made of graphite. In operation, the boat is loaded with a charge of polycrystalline gallium arsenide, with the arsenic kept at the other end of the tube.

As the furnace is moved toward the right, the melt cools at one end. Usually, there is a seed placed at the left end of the boat to establish a specific crystal orientation. The gradual freezing (solidification) of the melt allows a single crystal to propagate at the liquid-solid interface. Eventually, a single crystal of gallium arsenide is grown. The impurity distribution can be described essentially by Eqs. 9 and 15, where the growth rate is given by the traversing speed of the furnace.

TABLE 2 EQUILIBRIUM SEGREGATION COEFFICIENTS FOR DOPANTS IN GAAS

| Dopant | k_0 | Type |
|--------|-----------------------|----------------|
| Be | 3 | <i>p</i> |
| Mg | 0.1 | <i>p</i> |
| Zn | 4×10^{-1} | <i>p</i> |
| C | 0.8 | <i>n/p</i> |
| Si | 1.85×10^{-1} | <i>n/p</i> |
| Ge | 2.8×10^{-2} | <i>n/p</i> |
| S | 0.5 | <i>n</i> |
| Se | 5.0×10^{-1} | <i>n</i> |
| Sn | 5.2×10^{-2} | <i>n</i> |
| Te | 6.8×10^{-2} | <i>n</i> |
| Cr | 1.03×10^{-4} | Semiinsulating |
| Fe | 1.0×10^{-3} | Semiinsulating |

**Fig. 11 Bridgman technique for growing single-crystal gallium arsenide and a temperature profile of the furnace.**

► 11.4 MATERIAL CHARACTERIZATION

11.4.1 Wafer Shaping

After a crystal is grown, the first shaping operation is to remove the seed and the other end of the ingot, which is last to solidify.¹ The next operation is to grind the surface so that the diameter of the material is defined. After that, one or more flat regions are ground along the length of the ingot. These regions, or *flats*, mark the specific crystal orientation of the ingot and the conductivity type of the material. The largest flat, the *primary flat*, allows a mechanical locator in automatic processing equipment to position the wafer and to orient the devices relative

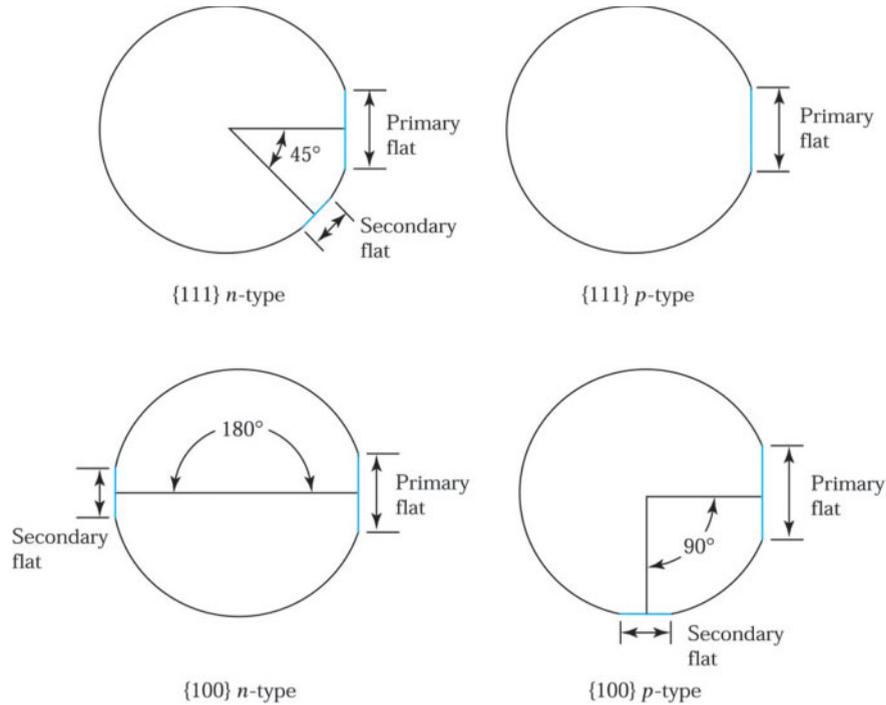


Fig. 12 Identifying flats on a semiconductor wafer.

to the crystal. Other smaller flats, called *secondary flats*, are ground to identify the orientation and conductivity type of the crystal, as shown in Fig. 12. For crystals with diameters equal or larger than 200 mm, no flats are ground. Instead, a small groove is ground along the length of the ingot.

The ingot is now ready to be sliced by diamond saw into wafers. Slicing determines four wafer parameters: *surface orientation* (e.g., $\langle 111 \rangle$ or $\langle 100 \rangle$), *thickness* (e.g., 0.5–0.8 mm, depending on wafer diameter); *taper*, which is the wafer thickness variations from one end to another; and *bow*, which is the surface curvature of the wafer, measured from the center of the wafer to its edge.

After slicing, both sides of the wafer are lapped using a mixture of Al_2O_3 and glycerine to produce a typical flatness uniformity within $2\ \mu\text{m}$. The lapping operation usually leaves the surface and edges of the wafer damaged and contaminated. The damaged and contaminated regions can be removed by chemical etching (see Chapter 12). The final step of wafer shaping is polishing. Its purpose is to provide a smooth, specular surface where device features can be defined by lithographic processes (see Chapter 12), Figure 13 shows a 300 mm silicon ingot and polished wafers. Table 3 shows the specifications for 125, 150, 200, 300, and 450 mm diameter polished silicon wafers from the Semiconductor Equipment and Materials Institute (SEMI). As mentioned previously, for large crystals (≥ 200 mm diameter) no flats are ground; instead, a groove is made on the edge of the wafer for positioning and orientation purposes.

Gallium arsenide is a more fragile material than silicon. Although the basic shaping operation of gallium arsenide is essentially the same as that for silicon, greater care must be exercised in gallium arsenide wafer preparation. The state of gallium arsenide technology is relatively primitive compared with that for silicon. However, the technology of group III-V compounds has advanced partly because of the advances in silicon technology.



Fig. 13 300 mm (12 in.) ingot and polished silicon wafers

TABLE 3 SPECIFICATIONS FOR POLISHED MONOCRYSTALLINE SILICON WAFERS

| Parameter | 125 mm | 150 mm | 200 mm | 300 mm | 450 mm |
|--------------------------------|------------|----------|-----------------|-------------|-----------|
| Diameter (mm) | 125±1 | 150±1 | 200±1 | 300±1 | 450±1 |
| Thickness (mm) | 0.6–0.65 | 0.65–0.7 | 0.715–0.735 | 0.755–0.775 | 0.78–0.80 |
| Primary flat length (mm) | 40–45 | 55–60 | NA ^a | NA | NA |
| Secondary flat length (mm) | 25–30 | 35–40 | NA | NA | NA |
| Bow (μm) | 70 | 60 | 30 | < 30 | < 30 |
| Total thickness variation (μm) | 65 | 50 | 10 | < 10 | < 10 |
| Surface orientation | (100) ± 1° | Same | Same | Same | Same |
| | (111) ± 1° | Same | Same | Same | Same |

^aNA: not available.

11.4.2 Crystal Characterization

Crystal Defects

A real crystal (such as a silicon wafer) differs from the ideal crystal in important ways. It is finite; thus, surface atoms are incompletely bonded. Furthermore, it has defects, which strongly influence the electrical, mechanical, and optical properties of the semiconductor. There are four categories of defects: point defects, line defects, area defects, and volume defects.

Figure 14 shows several forms of *point defects*.^{1,9} Any foreign atom incorporated into the lattice at either a substitutional site [i.e., at a regular lattice site (Fig. 14a)] or interstitial site [i.e., between regular lattice sites (Fig. 14b)] is a point defect. A missing atom in the lattice creates a vacancy, also considered a point defect (Fig. 14c). A host atom that is situated between regular lattice sites and adjacent to a vacancy is called a *Frenkel defect* (Fig. 14d). Point defects are particularly important subjects in the kinetics of diffusion and oxidation processes. These topics are considered in Chapters 12 and 14.

The next class of defects is the *line defect*, also called a dislocation.¹⁰ There are two types of dislocations: the edge and screw types. Figure 15a is a schematic representation of an edge dislocation in a cubic lattice: an extra plane of atoms *AB* is inserted into the lattice. The line of the dislocation would be perpendicular to the plane of the page.

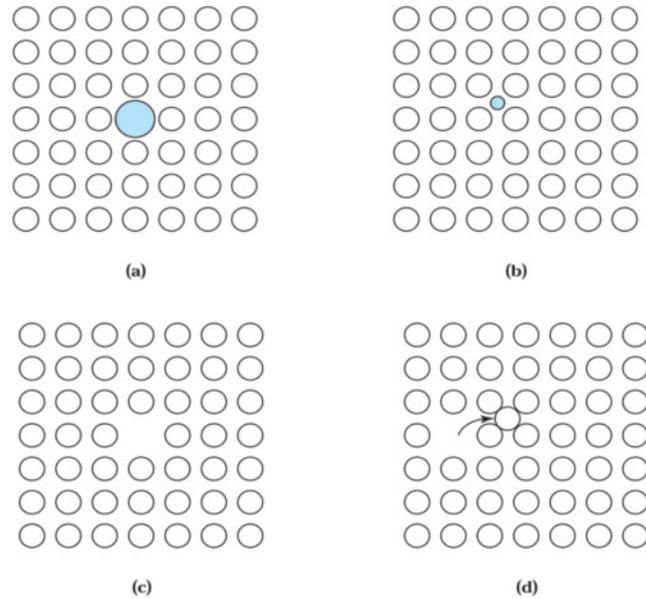


Fig. 14 Point defects. (a) Substitutional impurity. (b) Interstitial impurity. (c) Lattice vacancy. (d) Frenkel-type defect.⁹

The screw dislocation may be considered as being produced by cutting the crystal partway through and pushing the upper part one lattice spacing over, as shown in Fig. 15*b*. Line defects in devices are undesirable because they act as precipitation sites for metallic impurities, which may degrade device performance.

Area defects represent large area discontinuities in the lattice. Typical defects are twins and grain boundaries. Twinning represents a change in the crystal orientation across a plane. A grain boundary is a transition between crystals having no particular orientational relationship to one another. Such defects appear during crystal growth. Another area defect is the stacking fault.⁹ In this defect, the stacking sequence of atomic layer is interrupted. In Fig. 16 the sequence of atoms in a stack is *ABCABC* When a part of layer such as *C* in Fig. 16*a* is missing, it is called an intrinsic stacking fault. If an extra plane such as *A* in Fig. 16*b* is inserted between layers *B* and *C*, it is an extrinsic stacking fault. These defects may appear during crystal growth. Crystals having these area defects are not usable for integrated-circuit manufacture and are discarded.

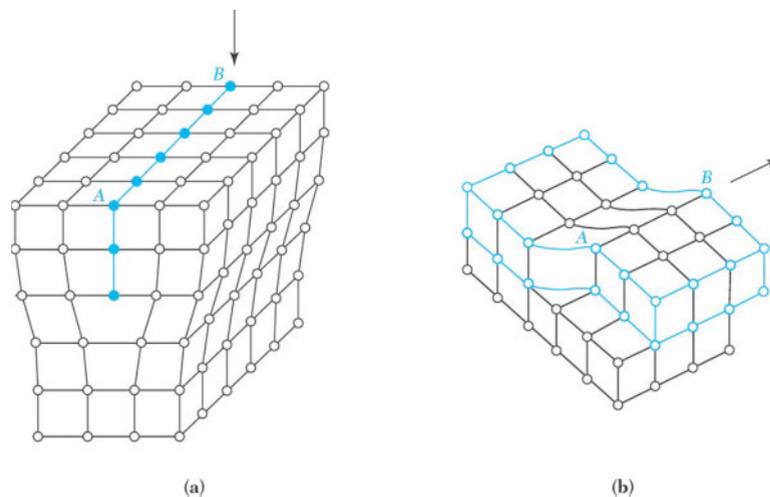


Fig. 15 (a) Edge and (b) screw dislocation formation in cubic crystals.¹⁰

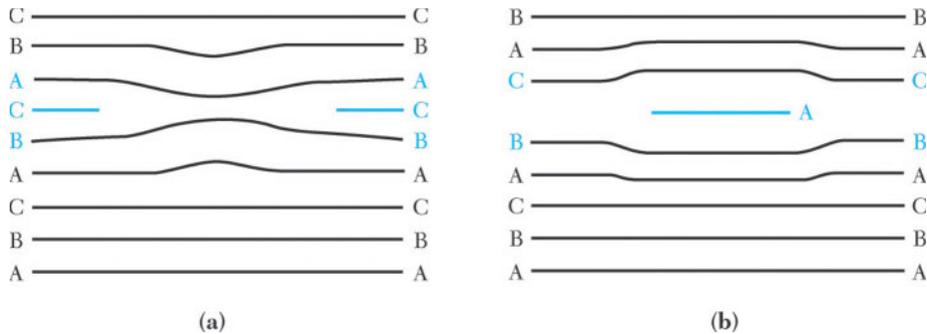


Fig. 16 Stacking faults in a semiconductor. (a) Intrinsic stacking fault. (b) Extrinsic stacking fault.⁹

Precipitates of impurities or dopant atoms make up the fourth class of defects, the *volume defects*. These defects arise because of the inherent solubility of the impurity in the host lattice. There is a specific concentration of impurity that the host lattice can accept in a solid solution of itself and the impurity. Figure 17 shows solubility versus temperature for a variety of elements in silicon.¹¹ The solubility of most impurities decreases with decreasing temperature. Thus, at a given temperature, if an impurity is introduced to the maximum concentration allowed by its solubility and the crystal is then cooled to a lower temperature, the crystal can achieve an equilibrium state only by precipitating the impurity atoms in excess of the solubility level. However, the volume mismatch between the host lattice and the precipitates results in dislocations.

Material Properties

Table 4 compares silicon characteristics and the requirements for ultra-large-scale integration* (ULSI).^{12,13} The semiconductor material properties listed in Table 4 can be measured by various methods. The resistivity is measured by the four-point probe method discussed in Section 2.1, Chapter 2, and the minority-carrier lifetime can be measured by the photoconductivity method considered in Section 2.3, Chapter 2. The trace impurities such as oxygen and carbon in silicon can be analyzed by the secondary-ion-mass spectroscopy (SIMS) techniques to be described in Chapter 14. Note that although current capabilities can meet most of the wafer specifications listed in Table 3, many improvements are needed to satisfy the stringent requirements for ULSI technology.¹³

The oxygen and carbon concentrations are substantially higher in Czochralski crystals than in float-zone crystals because of the dissolution of oxygen from the silica crucible and transport of carbon to the melt from the graphite susceptor during crystal growth. Typical carbon concentrations range from 10^{16} to about 10^{17} atoms/cm³, and carbon atoms in silicon occupy substitutional lattice sites. The presence of carbon is undesirable because it aids the formation of defects. Typical oxygen concentrations range from 10^{17} to 10^{18} atoms/cm³. Oxygen, however, has both deleterious and beneficial effects. It can act as a donor, distorting the resistivity of the crystal caused by intentional doping. However, oxygen in an interstitial lattice site can increase the yield strength of silicon. This beneficial effect increases with concentration until the oxygen begins to precipitate. Figure 17 shows that the typical oxygen concentration in Si wafer will precipitate at most common processing temperatures. A volume mismatch occurs as the precipitates grow in size, and results in a compressive strain on the lattice that is relieved by the formation of stacking faults (a type of dislocation) and other defects. The metal atoms do not fit in the silicon lattice easily because of their very different atomic sizes. They preferentially reside at sites in the silicon lattice where imperfections exist. Therefore, these defects can attract fast-diffusing metallic species [(diffusivities of metals are several orders of magnitude larger than those of common dopants like P, B, and As in silicon (Fig. 4 in Chapter 14)], which give rise to large junction leakage currents.

* The number of components in an ultralarge-scale integrated circuit is more than 10^7 .

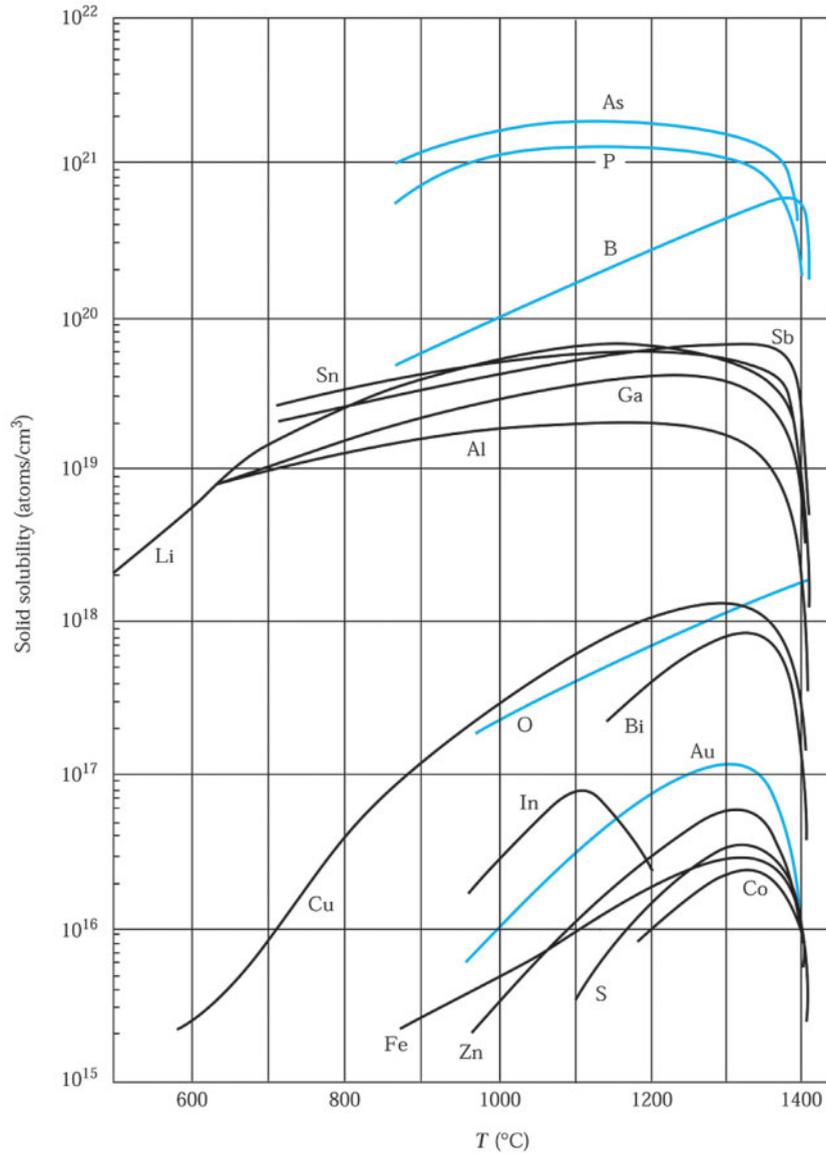


Fig. 17 Solid solubilities of impurity elements in silicon.¹¹

Certain precipitates can capture harmful impurities; this is called *gettering*. Gettering is a general term meaning a process that removes harmful impurities or defects from the region in a wafer where devices are fabricated. It is difficult to reduce the concentrations of impurities by purifying the wafers and by excluding metal contaminants from manufacturing environments. There are two basic methods for gettering. One is intrinsic gettering, which makes use of oxygen precipitates to getter metal atoms within the wafer bulk.

TABLE 4 COMPARISON OF SILICON MATERIAL CHARACTERISTICS AND REQUIREMENTS FOR ULSI

| Property ^a | Characteristics | | Requirements for ULSI |
|--|-----------------|--------------|------------------------|
| | Czochralski | Float zone | |
| Resistivity (phosphorus) <i>n</i> -type (ohm-cm) | 1–50 | 1–300 and up | 5–50 and up |
| Resistivity (antimony) <i>n</i> -type (ohm-cm) | 0.005–10 | — | 0.001–0.02 |
| Resistivity (boron) <i>p</i> -type (ohm-cm) | 0.005–50 | 1–300 | 5–50 and up |
| Resistivity gradient (four-point probe) (%) | 5–10 | 20 | < 1 |
| Minority carrier lifetime (μ s) | 30–300 | 50–500 | 300–1000 |
| Oxygen (ppma) | 5–25 | Not detected | Uniform and controlled |
| Carbon (ppma) | 1–5 | 0.1–1 | < 0.1 |
| Dislocation (before processing) (per cm^2) | ≤ 500 | ≤ 500 | ≤ 1 |
| Diameter (mm) | Up to 200 | Up to 100 | Up to 300 |
| Slice bow (μm) | ≤ 25 | ≤ 25 | < 5 |
| Slice taper (μm) | ≤ 15 | ≤ 15 | < 5 |
| Surface flatness (μm) | ≤ 5 | ≤ 5 | < 1 |
| Heavy-metal impurities (ppba) | ≤ 1 | ≤ 0.01 | < 0.001 |

^appma, parts per million atoms; ppba, parts per billion atoms.

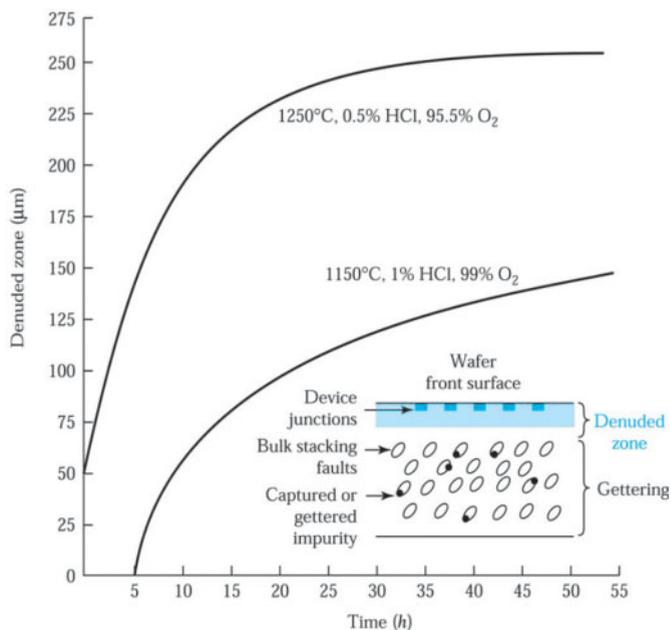


Fig. 18 Denuded zone width for two sets of processing conditions. Inset shows a schematic of the denuded zone and gettering sites in a wafer cross section.¹

The other, called extrinsic gettering, is created on the wafer backside by many methods, such as grinding, sandpaper abrasion, ion implantation, laser melting, depositing amorphous or polycrystalline films, high-concentration backside diffusion, etc. Once the backside damage sites are created, any subsequent high-temperature processing step will allow the metal atoms to diffuse to the backside where they can be trapped (note that metal atoms can easily diffuse completely through a silicon wafer at most common processing temperatures).

When the wafer is subjected to high-temperature treatment (e.g., 1050°C in N₂), oxygen evaporates from the surface. This lowers the oxygen content near the surface. The treatment creates a defect-free (or *denuded*) zone for device fabrication, as shown in the inset¹ of Fig. 18. Additional thermal cycles can be used to promote the formation of oxygen precipitates in the interior of the wafer for gettering of impurities. The depth of the defect-free zone depends on the time and temperature of the thermal cycle and on the diffusivity of oxygen in silicon. Measured results for the denuded zone are shown¹ in Fig. 18. It is possible to obtain Czochralski crystals of silicon that are virtually free of dislocations.

Commercial melt-grown materials of gallium arsenide are heavily contaminated by the crucible. However, for photonic applications, most requirements call for heavily doped materials (between 10¹⁷ and 10¹⁸ cm⁻³). For integrated circuits or for discrete MESFET (metal-semiconductor field-effect transistor) devices, undoped gallium arsenide can be used as the starting material, with a resistivity of 10⁹ Ω-cm. Oxygen is an undesirable impurity in GaAs because it can form a deep donor level, which contributes to a trapping charge in the bulk of the substrate and increases its resistivity. Oxygen contamination can be minimized by using graphite crucibles for melt growth. The dislocation content for Czochralski-grown gallium arsenide crystals is about two orders of magnitude higher than that of silicon. For Bridgman GaAs crystals, the dislocation density is about an order of magnitude lower than that of Czochralski-grown GaAs crystals.

► 11.5 EPITAXIAL-GROWTH TECHNIQUES

In an epitaxial process, the substrate wafer acts as the seed crystal. Epitaxial processes are differentiated from the melt-growth processes described in previous sections in that the epitaxial layer can be grown at a temperature substantially below the melting point, typically 30–50% lower. The common techniques for epitaxial growth are chemical-vapor deposition (CVD) and molecular-beam epitaxy (MBE).

11.5.1 Chemical-Vapor Deposition

CVD, also known as vapor-phase epitaxy (VPE), is a process whereby an epitaxial layer is formed by a chemical reaction between gaseous compounds. CVD can be performed at atmospheric pressure (APCVD) or at low pressure (LPCVD).

Figure 19 shows three common susceptors for epitaxial growth. Note that the geometric shape of the susceptor provides the name for the reactor: horizontal, pancake, and barrel susceptors—all made from graphite blocks. Susceptors in the epitaxial reactors are analogous to the crucible in the crystal-growing furnaces. Not only do they mechanically support the wafer, but in induction-heated reactors they also serve as the source of thermal energy for the reaction. The mechanism of CVD involves a number of steps: (a) the reactants such as the gases and dopants are transported to the substrate region, (b) they are transferred to the substrate surface where they are adsorbed, (c) a chemical reaction occurs, catalyzed at the surface, followed by growth of the epitaxial layer, (d) the gaseous products are desorbed into the main gas stream, and (e) the reaction products are transported out of the reaction chamber.

CVD for Silicon

Four silicon sources have been used for VPE growth: silicon tetrachloride (SiCl₄), dichlorosilane (SiH₂Cl₂), trichlorosilane (SiHCl₃), and silane (SiH₄). Silicon tetrachloride has been the most studied and has the widest industrial use. The typical reaction temperature is 1200°C. Other silicon sources are used because of lower reaction temperatures. The substitution of a hydrogen atom for each chlorine atom from silicon tetrachloride permits about a 50°C reduction in the reaction temperature. The overall reaction of silicon tetrachloride that results in the growth of silicon layers is



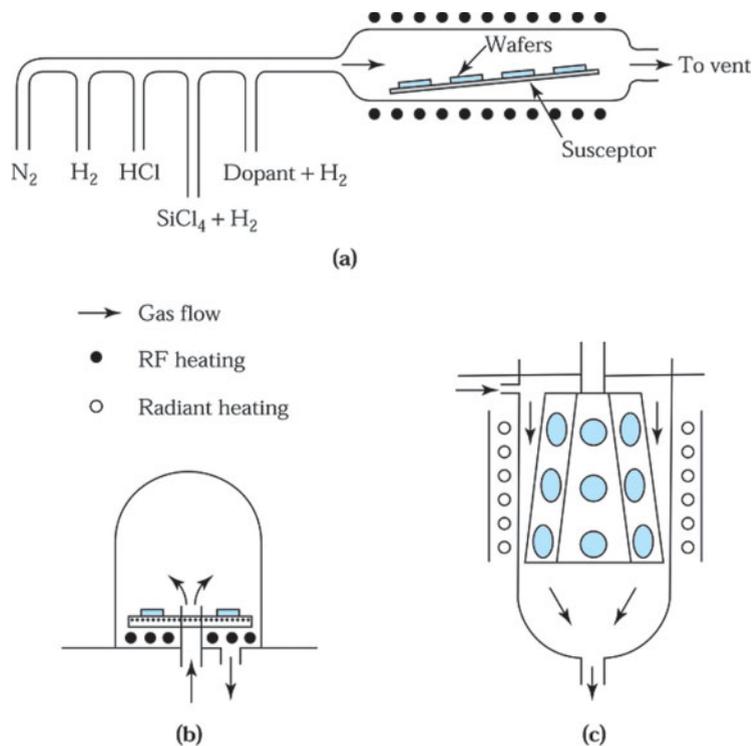
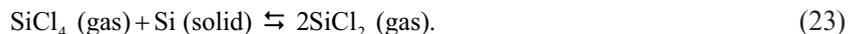


Fig. 19 Three common susceptors for chemical vapor deposition: (a) horizontal, (b) pancake, and (c) barrel susceptor.

An additional competing reaction is taking place along with that given in Eq. 22:



As a result, if the silicon tetrachloride concentration is too high, etching rather than growth of silicon will take place. Figure 20 shows the effect of the concentration of silicon tetrachloride in the gas on the reaction, where the *mole fraction* is defined as the ratio of the number of molecules of a given species to the total number of molecules.¹⁴ Note that initially the growth rate increases linearly with increasing concentration of silicon tetrachloride. As the concentration of silicon tetrachloride is increased, a maximum growth rate is reached. Beyond that, the growth rate starts to decrease and eventually etching of the silicon will occur. Silicon is usually grown in the low-concentration region, as indicated in Fig. 20.

The reaction of Eq. 22 is reversible, that is, it can take place in either direction. If the carrier gas entering the reactor contains hydrochloric acid, removal or etching will take place. Actually, this etching operation is used for in-situ cleaning of the silicon wafer and coating on the reactor chamber wall prior to epitaxial growth.

The dopant is introduced at the same time as the silicon tetrachloride during epitaxial growth (Fig. 19a). Gaseous diborane (B_2H_6) is used as the *p*-type dopant, whereas phosphine (PH_3) and arsine (AsH_3) are used as *n*-type dopants. Gas mixtures are ordinarily used with hydrogen as the diluent to allow reasonable control of flow rates for the desired doping concentration. The dopant chemistry for arsine is illustrated in Fig. 21, which shows arsine being adsorbed on the surface, decomposing, and being incorporated into the growing layer. Figure 21 also shows the growth mechanisms at the surface, which are based on the surface adsorption of host atoms (silicon) as well as the dopant atom (e.g., arsenic) and the movement of these atoms toward the ledge sites.¹⁵ To give these adsorbed atoms sufficient mobility for finding their proper positions within the crystal lattice, epitaxial growth needs relatively high temperatures.

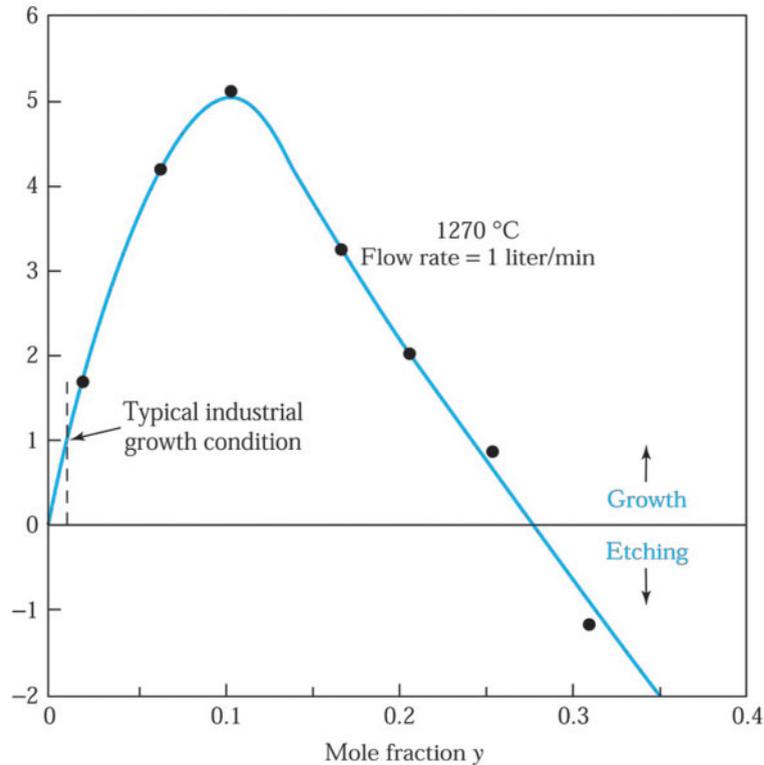


Fig. 20 Effect of SiCl_4 concentration on silicon epitaxial growth.¹⁴

CVD for GaAs

For gallium arsenide, the basic setup is similar to that shown in Fig. 19a. Since gallium arsenide decomposes into gallium and arsenic upon evaporation, its direct transport in the vapor phase is not possible. One approach is the use of As_4 for the arsenic component and gallium chloride (GaCl_3) for the gallium component. The overall reaction leading to epitaxial growth of gallium arsenide is



The As_4 is generated by thermal decomposition of arsine (AsH_3):



and the gallium chloride is generated by the reaction



The reactants are introduced into a reactor with a carrier gas (e.g., H_2). Usually, the temperature for Eq. 24b is 800°C . The growth temperature of GaAs epilayer for Eq. 24 is below 750°C . A two-zone reactor is needed for this epitaxial growth. Moreover, both reactions are exothermic: the epitaxy requires a reactor with hot walls. The reactions are near equilibrium condition, and process control is difficult. During the epitaxy, there must be sufficient arsenic overpressure to prevent thermal decomposition of the substrate and the growing layer.

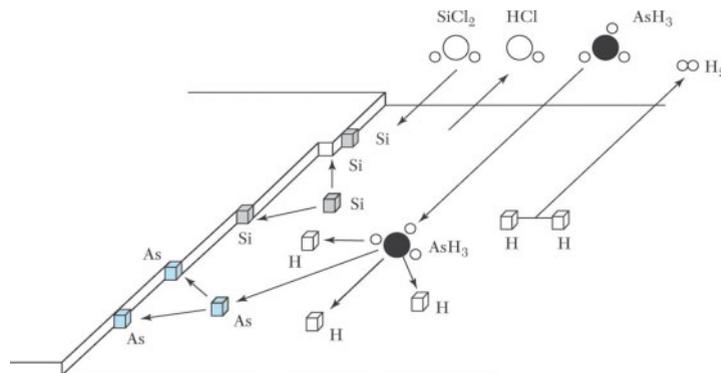
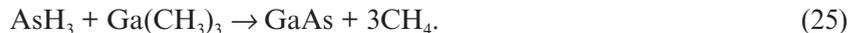


Fig. 21 Schematic representation of arsenic doping and the growing processes.¹⁵

Metalorganic CVD

Metalorganic CVD (MOCVD) is also a VPE process based on pyrolytic reactions. Unlike conventional CVD, MOCVD is distinguished by the chemical nature of the precursor. It is important for those elements that do not form stable hydrides or halides but that form stable metalorganic compounds with reasonable vapor pressure. MOCVD has been extensively applied in the heteroepitaxial growth of III-V and II-VI compounds.

To grow GaAs, we can use metalorganic compounds such as trimethylgallium $\text{Ga}(\text{CH}_3)_3$ for the gallium component and arsine AsH_3 for the arsenic component. Both chemicals can be transported in vapor form into the reactor. The overall reaction is



For Al-containing compounds, such as AlAs, we can use trimethylaluminum $\text{Al}(\text{CH}_3)_3$. During epitaxy, the GaAs is doped by introducing dopants in vapor form. Diethylzinc $\text{Zn}(\text{C}_2\text{H}_5)_2$ and diethylcadmium $\text{Cd}(\text{C}_2\text{H}_5)_2$ are typical *p*-type dopants and silane SiH_4 is an *n*-type dopant for III-V compounds. The hydrides of sulfur and selenium or tetramethyltin are also used for *n*-type dopants and chromyl chloride is used to dope chromium into GaAs to form semiinsulating layers. Since these compounds are highly poisonous and often spontaneously inflammable in air, rigorous safety precautions are necessary in the MOCVD process.

A schematic of an MOCVD reactor is shown¹⁶ in Fig. 22. Due to the endothermic reaction, a reactor with a cold wall is used. Typically, the metalorganic compound is transported to the quartz reaction vessel by hydrogen carrier gas, where it is mixed with AsH_3 in the case of GaAs growth. The chemical reaction is induced by heating the gases to $600^\circ\text{--}800^\circ\text{C}$ above a substrate placed on a graphite susceptor using radio-frequency heating. A pyrolytic reaction forms the GaAs layer. The advantages of using metalorganics are that they are volatile at moderately low temperatures and there are no troublesome liquid Ga or In sources in the reactor. A single hot zone and nonequilibrium (one-way) reaction make the control of MOCVD easier.

11.5.2 Molecular-Beam Epitaxy

MBE¹⁷ is an epitaxial process involving the reaction of one or more thermal beams of atoms or molecules with a crystalline surface under ultrahigh-vacuum conditions ($\sim 10^{-8}$ Pa).[§] MBE can achieve precise control in both chemical compositions and doping profiles. Single-crystal multilayer structures with dimensions on the order of atomic layers can be made using MBE. Thus, the MBE method enables the precise fabrication of semiconductor heterostructures having thin layers from a fraction of a micron down to a monolayer. In general, MBE growth rates are quite low, and for GaAs, a value of $1 \mu\text{m/hr}$ is typical.

[§]The international unit for pressure is the Pascal (Pa); $1 \text{ Pa} = 1 \text{ N/m}^2$. However, various other units have been used. The conversion of these units is: $1 \text{ atm} = 760 \text{ mm Hg} = 760 \text{ Torr} = 1.013 \times 10^5 \text{ Pa}$.

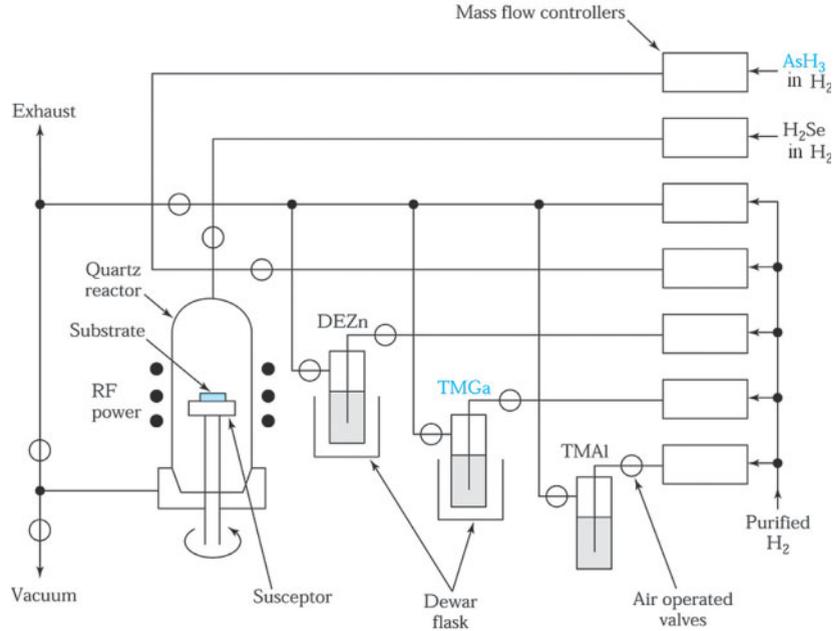


Fig. 22 Schematic diagram of a vertical atmospheric-pressure metalorganic chemical-vapor deposition (MOCVD) reactor.¹⁶ DEZn is diethylozinc $Zn(C_2H_5)_2$, TMGa is trimethylgallium $Ga(CH_3)_3$, and TMAI is trimethylaluminum $Al(CH_3)_3$.

Figure 23 shows a schematic of an MBE system for gallium arsenide and related III-V compounds such as $Al_xGa_{1-x}As$. The system represents the ultimate in film deposition control, cleanliness, and in-situ chemical characterization capability. Separate effusion ovens made of pyrolytic boron nitride are used for Ga, As, and the dopants. All the effusion ovens are housed in an ultrahigh-vacuum chamber ($\sim 10^{-8}$ Pa). The temperature of each oven is adjusted to give the desired evaporation rate. The substrate holder rotates continuously to achieve uniform epitaxial layers (e.g., $\pm 1\%$ in doping variations and $\pm 0.5\%$ in thickness variations).

To grow GaAs, an overpressure of As is maintained, since the sticking coefficient of Ga to GaAs is unity, whereas that for As is zero, unless there is a previously deposited Ga layer. For a silicon MBE system, an electron gun is used to evaporate silicon. One or more effusion ovens are used for the dopants. Effusion ovens behave like small-area sources and exhibit a $\cos\theta$ emission, where θ is the angle between the direction of the source and the normal to the substrate surface.

MBE uses an evaporation method in a vacuum system. An important parameter for vacuum technology is the molecular impingement rate, that is, how many molecules impinge on a unit area of the substrate per unit time. The impingement rate ϕ is a function of the molecular weight, temperature, and pressure. The rate is derived in Appendix K and can be expressed as¹⁸

$$\phi = P(2\pi mkT)^{-1/2} \quad (26)$$

or

$$\phi = 2.64 \times 10^{20} \left(\frac{P}{\sqrt{MT}} \right) \text{ molecules / cm}^2\text{-s,} \quad (26a)$$

where P is the pressure in Pa, m is the mass of a molecule in kg, k is Boltzmann's constant in J/K, T is the temperature in Kelvin, and M is the molecular weight. Therefore, at 300 K and 10^{-4} Pa pressure, the impingement rate is 2.7×10^{14} molecules/cm²-s for oxygen ($M = 32$).

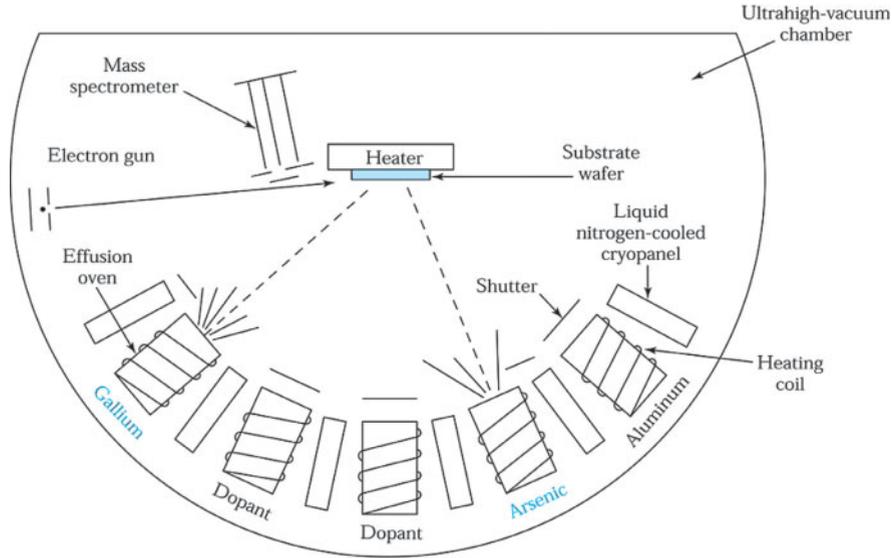


Fig. 23 Arrangement of the sources and substrate in a conventional molecular-beam epitaxy (MBE) system. (Courtesy of M. B. Panish, Bell Laboratories, Alcatel-Lucent Co.)

▶ EXAMPLE 3

At 300 K, the molecular diameter of oxygen is 3.64 \AA , and the number of molecules per unit area N_s is $7.54 \times 10^{14} \text{ cm}^{-2}$. Find the time required to form a monolayer of oxygen at pressures of 1 , 10^{-4} , and 10^{-8} Pa .

SOLUTION The time required to form a monolayer (assuming 100% sticking) is obtained from the impingement rate:

$$t = \frac{N_s}{\phi} = \frac{N_s \sqrt{MT}}{2.64 \times 10^{20} P}$$

Therefore,

$$\begin{aligned} t &= 2.8 \times 10^{-4} \approx 0.28 \text{ ms} && \text{at } 1 \text{ Pa,} \\ &= 2.8 \text{ s} && \text{at } 10^{-4} \text{ pa,} \\ &= 7.7 \text{ hr} && \text{at } 10^{-8} \text{ pa.} \end{aligned}$$

To avoid contamination of the epitaxial layer, it is of paramount importance to maintain ultrahigh-vacuum conditions ($\sim 10^{-8} \text{ Pa}$) for the MBE process. ◀

During molecular motion, molecules will collide with other molecules. The average distance traversed by all the molecules between successive collisions with each other is defined as the mean free path. It can be derived from a simple collision theory. A molecule having a diameter d and a velocity v will move a distance $v\delta t$ in the time δt . The molecule suffers a collision with another molecule if its center is anywhere within a distance d of the center of another molecule. Therefore, it sweeps out (without collision) a cylinder of diameter $2d$. The volume of the cylinder is

$$\delta V = \frac{\pi}{4} (2d)^2 v \delta t = \pi d^2 v \delta t. \quad (27)$$

Since there are n molecules/ cm^3 , the volume associated with one molecule is on the average $1/n \text{ cm}^3$. When the volume δV is equal to $1/n$, it must contain on the average one other molecule; thus, a collision would have occurred. Setting $\tau = \delta t$ as the average time between collision, we have

$$\frac{1}{n} = \pi d^2 v \tau, \quad (28)$$

and the mean free path λ is then

$$\lambda = v \tau = \frac{1}{\pi n d^2} = \frac{kT}{\pi P d^2}. \quad (29)$$

A more rigorous derivation gives

$$\lambda = \frac{kT}{\sqrt{2} \pi P d^2} \quad (30)$$

and

$$\lambda = \frac{0.66}{P(\text{in Pa})} \text{ cm} \quad (31)$$

for air molecules (equivalent molecular diameter of 3.7 \AA) at room temperature. Therefore, at a system pressure of 10^{-8} Pa , λ would be 660 km.

► EXAMPLE 4

Assume an effusion oven geometry of area $A = 5 \text{ cm}^2$ and a distance L between the top of the oven and the gallium arsenide substrate of 10 cm. Calculate the MBE growth rate for the effusion oven filled with gallium arsenide at 900°C . The surface density of gallium atom is $6 \times 10^{14} \text{ cm}^{-2}$, and the average thickness of a monolayer is 2.8 \AA .

SOLUTION

On heating gallium arsenide, the volatile arsenic vaporizes first, leaving a gallium-rich solution. Therefore, only the pressures marked Ga-rich in Fig. 10 are of interest. The pressure at 900°C is $5.5 \times 10^{-2} \text{ Pa}$ for gallium and 1.1 Pa for arsenic (As_2). The arrival rate can be obtained from the impingement rate (Eq. 26a) by multiplying it by $A/\pi L^2$:

$$\text{Arrival rate} = 2.64 \times 10^{20} \left(\frac{P}{\sqrt{MT}} \right) \left(\frac{A}{\pi L^2} \right) \text{ molecules / cm}^2\text{-s.}$$

The molecular weight M is 69.72 for Ga and 74.92×2 for As_2 . Substituting values of P , M , and T (1173 K) into the above equation gives

$$\begin{aligned} \text{Arrival rate} &= 8.2 \times 10^{14} / \text{cm}^2\text{-s} && \text{for Ga,} \\ &= 1.1 \times 10^{16} / \text{cm}^2\text{-s} && \text{for As}_2. \end{aligned}$$

The growth rate of gallium arsenide is found to be governed by the arrival rate of gallium. The growth rate is

$$\frac{8.2 \times 10^{14} \times 2.8}{6 \times 10^{14}} \approx 0.38 \text{ nm / s} = 23 \text{ nm / min.}$$

Note that the growth rate is relatively low compared with that of VPE. ◀

There are two ways to clean a surface in situ for MBE. High-temperature baking can decompose native oxide and remove other adsorbed species by evaporation or diffusion into the wafer. Another approach is to use a low-energy ion beam of an inert gas to sputter-clean the surface, followed by a low-temperature annealing to reorder the surface lattice structure.

MBE can use a wider variety of dopants than CVD and MOCVD, and the doping profile can be exactly controlled. However, the doping process is similar to the vapor-phase growth process: a flux of evaporated dopant atoms arrives at a favorable lattice site and is incorporated along the growing interface.

Fine control of the doping profile is achieved by adjusting the dopant flux relative to the flux of silicon atoms (for silicon epitaxial films) or gallium atoms (for gallium arsenide epitaxial films). It is also possible to dope the epitaxial film using a low-current, low-energy ion beam to implant the dopant (see Chapter 14).

The substrate temperatures for MBE range from 400°–900°C; and the growth rates range from 0.001 to 0.3 $\mu\text{m}/\text{min}$. Because of the low-temperature process and low growth rate, many unique doping profiles and alloy compositions not obtainable from conventional CVD can be produced in MBE. Many novel structures have been made using MBE, among them the *superlattice* and the heterojunction field-effect transistors discussed in Chapter 7.

A further development in MBE has replaced the group III elemental sources by metalorganic compounds such as trimethylgallium (TMG) or triethylgallium (TEG). This approach is called metalorganic molecular-beam epitaxy (MOMBE) and is also referred to as chemical-beam epitaxy (CBE). Although closely related to MOCVD, it is considered a special form of MBE. The metalorganics are sufficiently volatile that they can be admitted directly into the MBE growth chamber as a beam and are not decomposed before forming the beam. The dopants are generally elemental sources, typically Be for *p*-type and Si or Sn for *n*-type GaAs epitaxial layers.

► 11.6 STRUCTURES AND DEFECTS IN EPITAXIAL LAYERS

11.6.1 Lattice-Matched and Strained-Layer Epitaxy

For conventional homoepitaxial growth, a single-crystal semiconductor layer is grown on a single-crystal semiconductor substrate. The semiconductor layer and the substrate are the same material and have the same lattice constant. Therefore, homoepitaxy is, by definition, a lattice-matched epitaxial process. The homoepitaxial process offers an important means of controlling the doping profiles so that device and circuit performance can be optimized. For example, an *n*-type silicon layer with a relatively low doping concentration can be grown epitaxially on an n^+ -silicon substrate. This structure substantially reduces the series resistance associated with the substrate.

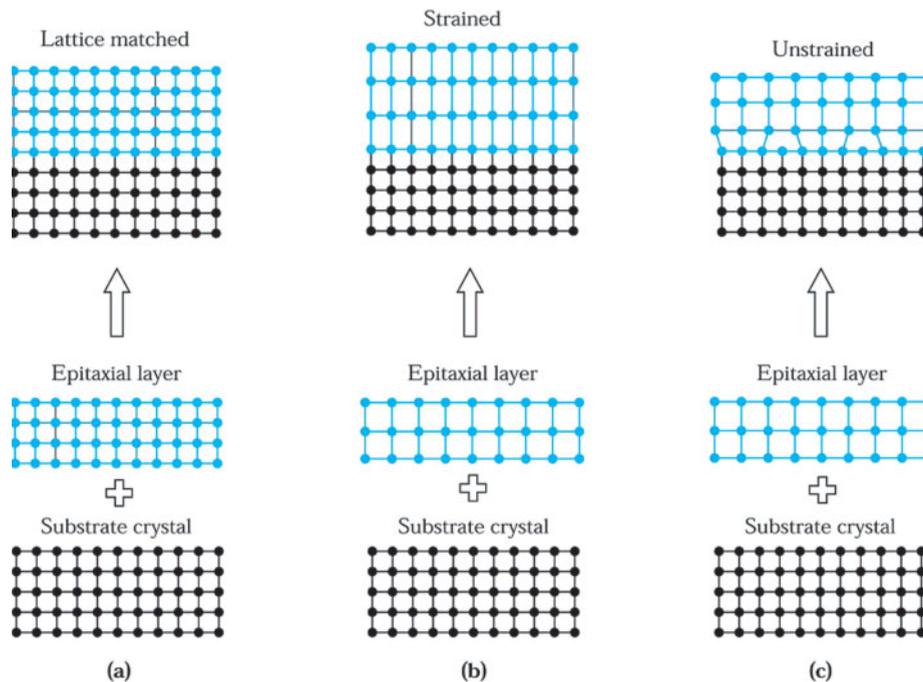


Fig. 24 Schematic of (a) lattice-matched, (b) strained, and (c) unstrained hetero-epitaxial structures.¹⁹ Homoepitaxy is structurally identical to lattice-matched heteroepitaxy.

For heteroepitaxy, the epitaxial layer and the substrate are two different semiconductors, and the epitaxial layer must be grown in such a way that an idealized interfacial structure is maintained. This implies that atomic bonding across the interface must be continuous without interruption. Therefore, the two semiconductors must either have the same lattice spacing or be able to deform to adopt a common spacing. These two cases are referred to as lattice-matched epitaxy and strained-layer epitaxy.

Figure 24a shows lattice-matched epitaxy, where the substrate and the film have the same lattice constant. An important example is the epitaxial growth of $\text{Al}_x\text{Ga}_{1-x}\text{As}$ on a GaAs substrate where for any x between 0 and 1, the lattice constant of $\text{Al}_x\text{Ga}_{1-x}\text{As}$ differs from that of GaAs by less than 0.13%.

For the lattice-mismatched case, if the epitaxial layer has a larger lattice constant and is flexible, it will be compressed in the plane of growth to conform to the substrate spacing. Elastic forces then compel it to dilate in a direction perpendicular to the interface. This type of structure is called strained-layer epitaxy and is illustrated¹⁹ in Fig. 24b. On the other hand, if the epitaxial layer has a smaller lattice constant, it will be dilated in the plane of growth and compressed in a direction perpendicular to the interface. In the above strained-layer epitaxy, as the strained-layer thickness increases, the total number of atoms under strain or the distorted atomic bonds grows, and at some point misfit dislocations are nucleated to relieve the homogeneous strain energy. This thickness is referred to as the *critical layer thickness* for the system. Figure 24c shows the case in which there are edge dislocations at the interface.

The critical layer thicknesses for two material systems are shown²⁰ in Fig. 25. The upper curve is for the strained-layer epitaxy of a $\text{Ge}_x\text{Si}_{1-x}$ layer on a silicon substrate, and the lower curve is for a $\text{Ga}_{1-x}\text{In}_x\text{As}$ layer on a GaAs substrate. For example, for $\text{Ge}_{0.3}\text{Si}_{0.7}$ on silicon, the maximum epitaxial thickness is about 70 nm. For thicker films, edge dislocations will occur.

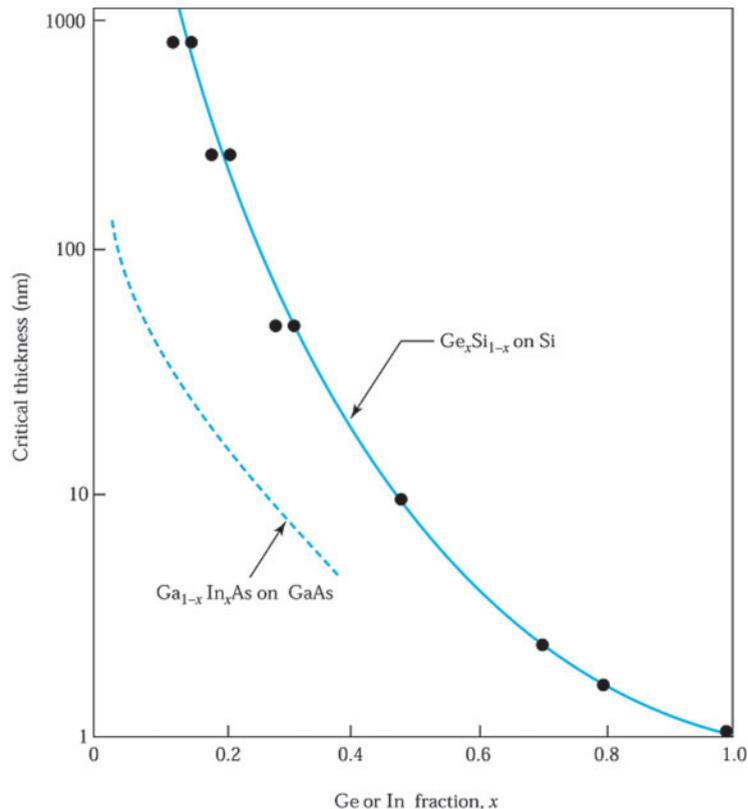


Fig. 25 Experimentally determined critical layer thickness for defect-free strained-layer epitaxy²⁰ of $\text{Ge}_x\text{Si}_{1-x}$ on Si, and $\text{Ga}_{1-x}\text{In}_x\text{As}$ on GaAs.

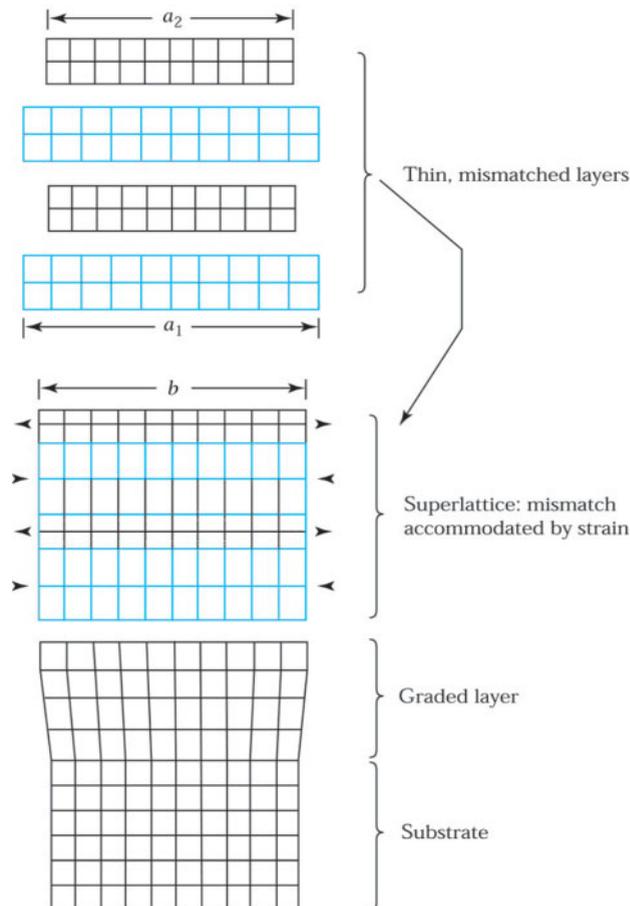


Fig. 26 Illustration of the elements and formation of a strained-layer superlattice.¹⁷ Arrows show the direction of the strain.

A related heteroepitaxial structure is the strained-layer superlattice (SLS). A superlattice is an artificial one-dimensional periodic structure constituted by different materials with a period of about 10 nm. Figure 26 shows¹⁷ a SLS having two semiconductors with different equilibrium lattice constants $a_1 > a_2$ grown in a structure with a common inplane lattice constant b , where $a_1 > b > a_2$. For sufficiently thin layers, the lattice mismatch is accommodated by uniform strains in the layers. Under these conditions, no misfit dislocations are generated at the interfaces, so high-quality crystalline materials can be obtained. These artificially structured materials, which can be grown by MBE, provide a new area in semiconductor research and permit new solid-state devices, especially for high-speed and photonic applications.

11.6.2 Compound Semiconductors on Silicon

The technologies of strained-layer, high- k gate oxide, nanowire, and multigate have been utilized in IC fabrication. To continue the development of IC industry, the heteroepitaxial technology has gained much attention in recent years. The superior properties (higher electron mobility, larger bandgap) of III-V compounds (e.g., GaAs, InP) make heteroepitaxial layers of these compounds on Si substrates attractive. In addition, this heteroepitaxial approach can provide lower cost, higher mechanical strength, better thermal conductivity, larger wafer area, and the possibility of monolithic integration of electronic and optical devices.

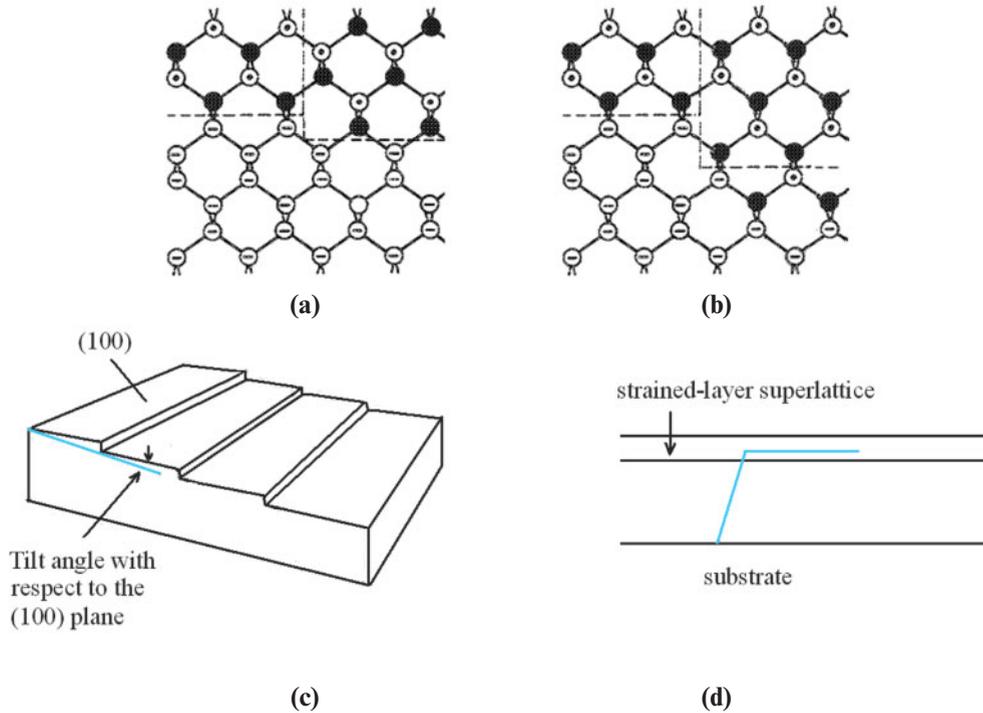


Fig. 27 (a) A single step leads to the formation of APD for a polar semiconductor grown on nonpolar semiconductor. (b) Double steps to eliminate APD. (c) Tilted substrate creates plateau to reduce dislocations. (d) Strong strain field bends the lattice-mismatch dislocation.

There are three problems in the heteroepitaxy process²¹: (1) antiphase domains (APD) from polar semiconductor grown on nonpolar semiconductor shown in Fig. 27a, (2) high dislocation density from large lattice mismatch (4% for GaAs/Si, 8% for InP/Si), and (3) cracking, bowing, and bending from large thermal-expansion-coefficient mismatch (GaAs is 2.2 times of Si) during cooling from the growth temperature.

There are many techniques used to eliminate or minimize these problems. The most common techniques are the tilted substrates and the pseudomorphic (strained-layer) superlattice buffer layer. The tilted substrate can create double steps to reduce antiphase domains, as shown in Fig. 27b. For GaAs/Si, 4% lattice mismatch means one dislocation generation for every 25 atomic planes. A substrate with a suitable tilted angle can create a plateau with a width less than 25 atoms on the substrate surface, as shown in Fig. 27c, and the lattice misfit dislocation can be reduced. The pseudomorphic superlattice buffer layer provides a strain field to bend the dislocation and prevent it from propagating into the active region of a device, as shown in Fig. 27d. However, there is no effective way to avoid thermal-expansion-coefficient mismatch. Although the stress during cooling is not large enough to cause dislocations, the repeated thermal cycles would cause bowing, bending, and even cracking. However, high-quality enhancement-mode $\text{In}_{0.7}\text{Ga}_{0.3}\text{As}$ quantum-well transistors on silicon substrate have been obtained by these techniques.²²

Recently, the critical thickness of a nanowire has been found to be roughly an order of magnitude larger than the critical thickness on the same substrate in thin film systems.²³ Nanowire heterostructures may exhibit defect-free interfaces even for systems with large lattice mismatch. This approach may provide another possible solution for the heteroepitaxial growth of III-V compounds on Si.

► SUMMARY

Several techniques are available to grow single crystals of silicon and gallium arsenide. For silicon crystals, we use sand (SiO_2) to produce polycrystalline silicon, which then serves as the raw material in a Czochralski puller. A seed crystal with the desired orientation is used to grow a large ingot from the melt. Over 90% of silicon crystals are prepared by this technique. During crystal growth, the dopant in the crystal will redistribute. A key parameter is the segregation coefficient, i.e., the ratio of the dopant concentration in the solid to that in the melt. Since most coefficients are less than 1, the melt becomes progressively enriched with the dopant as the crystal grows.

Another growth technique for silicon is the float-zone process. It offers lower contamination than that normally obtained from the Czochralski technique. Float-zone crystals are used mainly for high-power, high-voltage devices where high-resistivity materials are required.

To make GaAs, we use chemically pure gallium and arsenic as the starting materials that are synthesized to form polycrystalline GaAs. Single crystals of GaAs can be grown by the Czochralski technique. However, a liquid encapsulant (e.g., B_2O_3) is required to prevent decomposition of GaAs at the growth temperature. Another technique is the Bridgman process, which uses a two-zone furnace for gradual solidification of the melt.

After a crystal is grown, it usually goes through wafer-shaping operations to give an end product of highly polished wafers with a specified diameter, thickness, and surface orientation. For example, 300 mm silicon wafers for a MOSFET (metal-oxide-semiconductor field-effect transistor) fabrication line should have a diameter of 300 ± 1 mm, a thickness of 0.765 ± 0.01 mm, and a surface orientation of $(100) \pm 1^\circ$. Wafers with diameters larger than 300 mm are being manufactured for future integrated circuits. Their specifications are listed in Table 3.

A real crystal has defects that influence the electrical, mechanical, and optical properties of the semiconductor. These defects are point defects, line defects, area defects, and volume defects. We also discussed means to minimize such defects. For the more demanding ULSI applications, the dislocation density must be less than 1 per square centimeter. Other important requirements are listed in Table 4.

A technology closely related to crystal growth is the epitaxial process. In this process, the substrate wafer is the seed. High-quality, single-crystal films can be grown at temperatures 30%–50% lower than the melting point. The common techniques for epitaxial growth are chemical-vapor deposition (CVD), metalorganic CVD (MOCVD), and molecular-beam epitaxy (MBE). CVD and MOCVD are chemical deposition processes. Gases and dopants are transported in vapor form to the substrate, where a chemical reaction occurs that results in the deposition of the epitaxial layer. Inorganic compounds are used for CVD, whereas metalorganic compounds are used for MOCVD. MBE, on the other hand, is a physical deposition process. It is done by the evaporation of a species in an ultrahigh vacuum system. Because it is a low-temperature process that has a low growth rate, MBE can grow single-crystal, multilayer structures with dimensions on the order of atomic layers.

In addition to conventional homoepitaxy, such as n -type silicon on an n^+ -silicon substrate, we have also considered heteroepitaxy, which includes lattice-matched and strained-layer structures. For strained-layer epitaxy, there is a critical layer thickness above which edge dislocations will nucleate to relieve the strain energy.

The heteroepitaxy and superior properties of III-V compounds on Si are attractive in continuing development in the IC industry. Various means have been presented to minimize or even eliminate the problems. But they are not solved completely. The critical thickness of a nanowire has been found to be roughly an order of magnitude larger than the critical thickness on the same substrate in thin-film systems. This approach may provide another possible solution for the heteroepitaxial growth of III-V compounds on Si.

► REFERENCES

1. R. Doering and Y. Nishi, “*Handbook of Semiconductor Manufacturing Technology*,” 2nd Ed., CRC Press, FL, 2008.
2. C. W. Pearce, “Crystal Growth and Wafer Preparation” and “Epitaxy,” in S. M. Sze, Ed., *VLSI Technology*, McGraw-Hill, New York, 1983.
3. W. R. Runyan, *Silicon Semiconductor Technology*, McGraw-Hill, New York, 1965.

4. W. G. Pfann, *Zone Melting*, 2nd Ed., Wiley, New York, 1966.
5. E. W. Hass and M. S. Schnoller, "Phosphorus Doping of Silicon by Means of Neutron Irradiation," *IEEE Trans. Electron Devices*, **ED-23**, 803 (1976).
6. M. Hansen, *Constitution of Binary Alloys*, McGraw-Hill, New York, 1958.
7. S. K. Ghandhi, *VLSI Fabrication Principles*, Wiley, New York, 1983.
8. J. R. Arthur, "Vapor Pressures and Phase Equilibria in the GaAs System," *J. Phys. Chem. Solids*, **28**, 2257 (1967).
9. B. El-Kareh, *Fundamentals of Semiconductor Processing Technology*, Kluwer Academic, Boston, 1995.
10. C. A. Wert and R. M. Thomson, *Physics of Solids*, McGraw-Hill, New York, 1964.
11. (a) F. A. Trumbore, "Solid Solubilities of Impurity Elements in Germanium and Silicon," *Bell Syst. Tech. J.*, **39**, 205 (1960); (b) R. Hull, *Properties of Crystalline Silicon*, INSPEC, London, 1999.
12. Y. Matsushita, "Trend of Silicon Substrate Technologies for 0.25 μm Devices," *Proc. VLSI Technol. Workshop*, Honolulu (1996).
13. *The International Technology Roadmap for Semiconductors*, Semiconductor Industry Association, San Jose, CA, 2009.
14. A. S. Grove, *Physics and Technology of Semiconductor Devices*, Wiley, New York, 1967.
15. R. Reif, T. I. Kamins, and K. C. Saraswat, "A Model for Dopant Incorporation into Growing Silicon Epitaxial Films," *J. Electrochem. Soc.*, **126**, 644, 653 (1979).
16. R. D Dupuis, *Science*, "Metalorganic Chemical Vapor Deposition of III–V Semiconductors", **226**, 623 (1984).
17. M. A. Herman and H. Sitter, *Molecular Beam Epitaxy*, Springer-Verlag, Berlin, 1996.
18. A. Roth, *Vacuum Technology*, North-Holland, Amsterdam, 1976.
19. M. Ohring, *The Materials Science of Thin Films*, Academic, New York, 1992.
20. J. C. Bean, "The Growth of Novel Silicon Materials," *Physics Today*, **39**, 10, 36 (1986).
21. S. F. Fang, K. Adomi, S. Iyer, H. Morkoc, and H. Zabel, "Gallium arsenide and other compound semiconductors on silicon," *J. Appl. Phys.*, **68**, R3 (1990).
22. M. K. Hudait et al., "Heterogeneous integration of enhancement mode $\text{In}_{0.7}\text{Ga}_{0.3}\text{As}$ quantum well transistor on silicon substrate using thin (2 μm) composite buffer architecture for high-speed and low-voltage (0.5V) logic applications," *IEDM Tech. Dig.*, 625, 2007.
23. E. Ertekin, P. A. Greaney, and D. C. Chrzan, "Equilibrium limits of coherency in strained nanowire heterostructures," *J. Appl. Phys.*, **97**, 11, 114 325 (2005).

► **PROBLEMS (* DENOTES DIFFICULT PROBLEMS)**

FOR SECTION 10.1 SILICON CRYSTAL GROWTH FROM THE MELT

1. Plot the doping distribution of arsenic at distances of 10, 20, 30, 40, and 45 cm from the seed in a silicon ingot 50 cm long that has been pulled from a melt with an initial doping concentration of 10^{17} cm^{-3} .
2. In silicon, the lattice constant is 5.43 Å. Assume a hard-sphere model. (a) Calculate the radius of a silicon atom. (b) Determine the density of silicon atoms in atoms/cm³. (c) Use the Avogadro number to find the density of silicon.

3. Assuming that a 10 kg pure silicon charge is used, what amount of boron must be added to get boron-doped silicon having a resistivity of $0.01 \Omega\text{-cm}$ when one half of the ingot is grown?
4. A silicon wafer 1-mm thick having a diameter of 200 mm contains 5.41 mg of boron uniformly distributed in substitutional sites. Find (a) the boron concentration in atoms/cm³ and, (b) the average distance between boron atoms.
5. The seed crystal used in Czochralski processing is usually necked down to a small diameter (5.5 mm) as a means to initiate dislocation-free growth. If the critical yield strength of silicon is $2 \times 10^6 \text{ g/cm}^2$, calculate the maximum length of a silicon ingot 200 mm in diameter that can be supported by such a seed.
6. Plot the curve of C_s/C_0 value for $k_0 = 0.05$ in the Czochralski technique.
7. A Czochralski grown crystal is doped with boron. Why is the boron concentration larger at the tail end of the crystal than at the seed end?
8. Why is the impurity concentration larger in the center of the wafer than at its perimeter?

FOR SECTION 10.2 SILICON FLOAT-ZONE PROCESS

9. We use the float-zone process to purify a silicon ingot that contains a uniform gallium concentration of $5 \times 10^{16} \text{ cm}^{-3}$. One pass is made with a molten zone 2 cm long. Over what distance is the resulting gallium concentration below $5 \times 10^{15} \text{ cm}^{-3}$?
10. From Eq. 18 find the C_s/C_0 value at $x/L = 1$ and 2 with $k_c = 0.3$.
11. If p^+n abrupt-junction diodes are fabricated using the silicon materials shown in Fig. 8, find the percentage change of breakdown voltages for the conventionally doped silicon and the neutron-irradiated silicon.

FOR SECTION 10.3 GaAs CRYSTAL-GROWTH TECHNIQUES

12. From Fig. 9, if C_m is 20%, what fraction of the liquid at T_b will remain?
13. From Fig. 10, explain why the GaAs liquid always becomes gallium rich?

FOR SECTION 10.4 MATERIAL CHARACTERIZATION

14. The equilibrium density of vacancies n_s is given by $N \exp(-E_s/kT)$, where N is the density of semiconductor atoms and E_s is the energy of formation. Calculate n_s in silicon at 27°C, 900°C, and 1200°C. Assume $E_s = 2.3 \text{ eV}$.
15. Assume the energy of formation (E_f) of a Frenkel-type defect to be 1.1 eV and estimate the defect density at 27°C and 900°C. The equilibrium density of Frenkel type defects is given by $n_f = \sqrt{NN'} e^{-E_f/2kT}$, where N is the atomic density of silicon (cm⁻³) and N' is the density of available interstitial sites (cm⁻³), is represented by $N' = 1 \times 10^{27} e^{-3.8(\text{eV})/kT} \text{ cm}^{-3}$.
16. How many chips of area $400 \mu\text{m}^2$ can be placed on a wafer 300 mm in diameter? Explain your assumptions on the chip shape and unused wafer perimeter.

FOR SECTION 10.5 EPITAXIAL-GROWTH TECHNIQUES

- *17. Find the average molecular velocity of air at 300 K (the molecular weight of air is 29).
18. The distance between source and wafer in a deposition chamber is 15 cm. Estimate the pressure at which this distance becomes 10% of the mean free path of source molecules.

- *19. Find the number of atoms per unit area, N_s , needed to form a monolayer under close-packing conditions (i.e., each atom is in contact with its six neighboring atoms), assuming the diameter d of the atom is 4.68 \AA .
- *20. Assume an effusion oven geometry of $A = 5 \text{ cm}^2$ and $L = 12 \text{ cm}$. (a) Calculate the arrival rate of gallium and the MBE growth rate for the effusion oven filled with gallium arsenide at 970°C . (b) For a tin effusion oven operated at 700°C under the same geometry, calculate the doping concentration (assuming tin atoms are fully incorporated in the gallium arsenide grown at the aforementioned rate). The molecular weight of tin is 118.69 and the pressure for tin is $2.66 \times 10^{-6} \text{ Pa}$ at 700°C .

FOR SECTION 10.6 STRUCTURES AND DEFECTS IN EPITAXIAL LAYERS

21. Find the maximum percentage of In, i.e., the x value for $\text{Ga}_x\text{In}_{1-x}\text{As}$ film grown on GaAs substrate without formation of misfit dislocation, if the final film thickness is 10 nm.
22. The lattice misfit, f , of a film is defined as $f \equiv [a_0(s) - a_0(f)]/a_0(f) = \Delta a_0/a_0$, where $a_0(s)$ and $a_0(f)$ are the unstrained lattice constants of the substrate and film, respectively. Find the f values for InAs-GaAs and Ge-Si systems.

Film Formation

- ▶ 12.1 THERMAL OXIDATION
 - ▶ 12.2 CHEMICAL VAPOR DEPOSITION OF DIELECTRICS
 - ▶ 12.3 CHEMICAL VAPOR DEPOSITION OF POLYSILICON
 - ▶ 12.4 ATOMIC LAYER DEPOSITION
 - ▶ 12.5 METALLIZATION
 - ▶ SUMMARY
-

To fabricate discrete devices and integrated circuits, we use many different kinds of thin films. We can classify thin films into four groups: thermal oxides, dielectric layers, polycrystalline silicon, and metal films. Figure 1 shows a schematic view of a conventional silicon n -channel MOSFET (metal-oxide-semiconductor field-effect transistor) that uses all four groups of films. The first important thin film from the thermal oxide group is the gate-oxide layer, under which a conducting channel can be formed between the source and the drain. A related layer is the field oxide, which provides isolation from other devices. Both gate and field oxides generally are grown by a thermal oxidation process because only thermal oxidation can provide the highest-quality oxides having the lowest interface trap densities.

Dielectric layers such as silicon dioxide and silicon nitride are used for insulation between conducting layers, for diffusion and ion implantation masks, for capping doped films to prevent the loss of dopants, and for passivation to protect devices from impurities, moisture, and scratches. Polycrystalline silicon, usually referred to as polysilicon, is used as a gate electrode material in MOS devices, a conductive material for multilevel metallization, and a contact material for devices with shallow junctions. Metal films such as copper and silicides are used to form low-resistance interconnections, ohmic contacts, and rectifying metal-semiconductor barriers.

Specifically, we cover the following topics:

- The current density equation and its drift and diffusion components.
- The thermal oxidation process to form silicon dioxide (SiO_2).
- Chemical –vapor deposition techniques to form dielectrics and polysilicon films.
- Metallization and related global planarization.
- Atomic layer deposition to form thin films of the order of a monolayer.
- Characteristics of these thin films and their compatibility with integrated-circuit processing.

▶ 12.1 THERMAL OXIDATION

Semiconductors can be oxidized by various methods. These include thermal oxidation, electrochemical anodization, and plasma reaction. Among these methods, thermal oxidation is by far the most important for silicon devices. It is the key process in modern silicon integrated-circuit technology. For gallium arsenide, however, thermal oxidation results in generally nonstoichiometric films. The oxides provide poor electrical insulation and

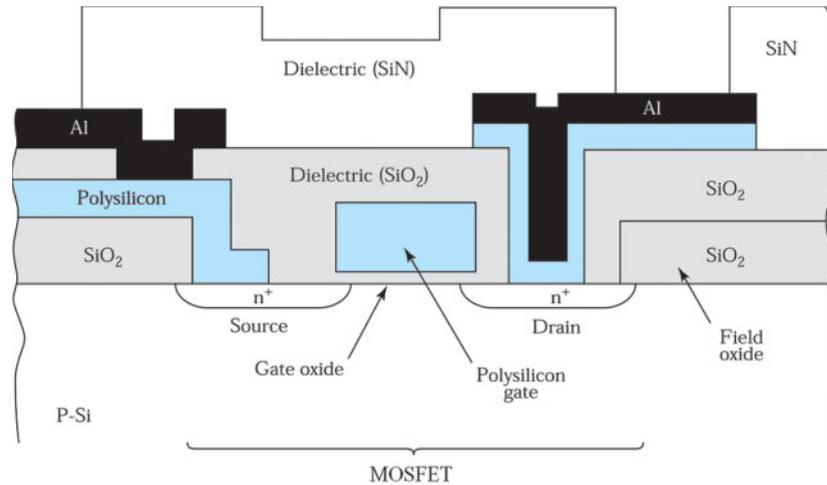


Fig. 1 Schematic cross section of a metal-oxide-semiconductor field-effect transistor (MOSFET).

semiconductor surface protection; hence, these oxides are rarely used in gallium arsenide technology. Consequently, in this section we concentrate on the thermal oxidation of silicon.

The basic thermal oxidation setup is shown¹ in Fig. 2. The reactor consists of a resistance-heated furnace, a cylindrical fused-quartz tube containing the silicon wafers held vertically in a slotted quartz boat, and a source of either pure dry oxygen or pure water vapor. The loading end of the furnace tube protrudes into a vertical flow hood where a filtered flow of air is maintained. Flow is directed as shown by the arrow in Fig. 2. The hood reduces dust and particulate matters in the air surrounding the wafers and minimizes contamination during wafer loading. The oxidation temperature is generally in the range of 900°-1200 °C and the typical gas flow rate is about 1 liter/min. The oxidation system uses microprocessors to regulate the gas flow sequence, to control the automatic insertion and removal of silicon wafers, to ramp the temperature up (i.e., to increase the furnace temperature linearly) from a low temperature to the oxidation temperature so that the wafers will not warp due to sudden temperature change, to maintain the oxidation temperature to within ±1°C, and to ramp the temperature down when oxidation is completed.

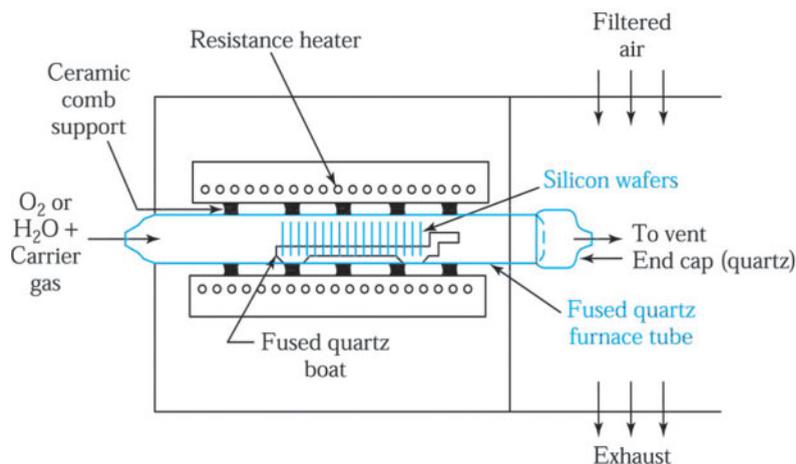
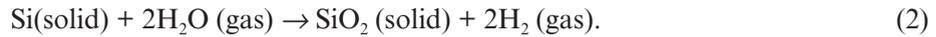
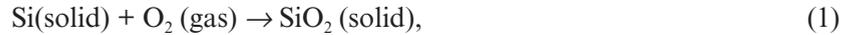


Fig. 2 Schematic cross section of a resistance-heated oxidation furnace.

12.1.1 Kinetics of Growth

The following chemical reactions describe the thermal oxidation of silicon in oxygen or water vapor:



The silicon-silicon dioxide interface moves into the silicon during the oxidation process. This creates a fresh interface region with surface contamination on the original silicon ending up on the oxide surface. The densities and molecular weights of silicon and silicon dioxide are used in the following example to show that growing an oxide of thickness x consumes a layer of silicon $0.44x$ thick (Fig. 3).

▶ EXAMPLE 1

If a silicon oxide layer of thickness x is grown by thermal oxidation, what is the thickness of silicon being consumed? The molecular weight of Si is 28.9 g/mol, and the density of Si is 2.33 g/cm³. The corresponding values for SiO₂ are 60.08 g/mol and 2.21 g/cm³.

SOLUTION The volume of 1 mol of silicon is

$$\frac{\text{Molecular weight of Si}}{\text{Density of Si}} = \frac{28.9 \text{ g / mole}}{2.33 \text{ g / cm}^3} = 12.06 \text{ cm}^3 / \text{mol}.$$

The volume of 1 mol of silicon dioxide is

$$\frac{\text{Molecular weight of SiO}_2}{\text{Density of SiO}_2} = \frac{60.08 \text{ g / mol}}{2.21 \text{ g / cm}^3} = 27.18 \text{ cm}^3 / \text{mol}.$$

Since 1 mol of silicon is converted to 1 mol of silicon dioxide,

$$\frac{\text{Thickness of Si} \times \text{area}}{\text{Thickness of SiO}_2 \times \text{area}} = \frac{\text{volume of 1 mol of Si}}{\text{volume of 1 mol of SiO}_2},$$

$$\frac{\text{Thickness of Si}}{\text{Thickness of SiO}_2} = \frac{12.06}{27.18} = 0.44,$$

Thickness of silicon = 0.44 (thickness of SiO₂).

For example, to grow a silicon dioxide layer of 100 nm, a layer of 44 nm of silicon is consumed. ◀

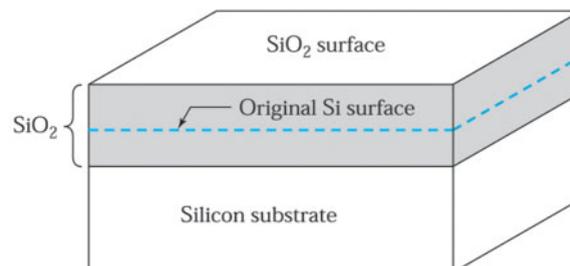


Fig. 3 Growth of silicon dioxide by thermal oxidation.

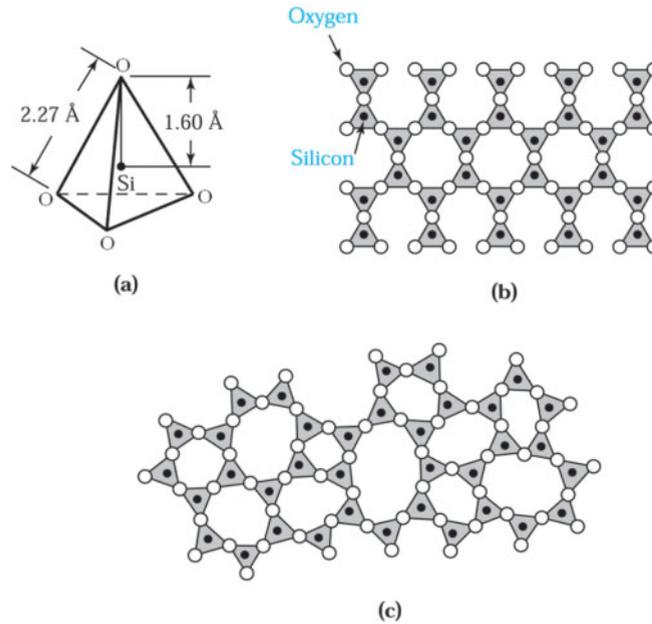


Fig. 4 (a) Basic structural unit of silicon dioxide. (b) Two-dimensional representation of a quartz crystal lattice. (c) Two-dimensional representation of the amorphous structure of silicon dioxide.¹

The basic structural unit of thermally grown silicon dioxide is a silicon atom surrounded tetrahedrally by four oxygen atoms, as illustrated¹ in Fig. 4a. The silicon-to-oxygen internuclear distance is 1.6 Å, and the oxygen-to-oxygen internuclear distance is 2.27 Å. These tetrahedra are joined together at their corners by oxygen bridges in a variety of ways to form the various phases or structures of silicon dioxide (also called silica). Silica has several crystalline structures (e.g., quartz) and an amorphous structure. When silicon is thermally oxidized, the silicon dioxide structure is amorphous. Typically amorphous silica has a density of 2.21 g/cm³ compared with 2.65 g/cm³ for quartz.

The basic difference between the crystalline and amorphous structures is that the former is a periodic structure, extending over many molecules, whereas the latter has no periodic structure at all. Figure 4b is a two-dimensional schematic diagram of a quartz crystalline structure made up of rings with six silicon atoms. Figure 4c is a two-dimensional schematic diagram of an amorphous structure for comparison. In the amorphous structure there is still a tendency to form characteristic rings with six silicon atoms. Note that the amorphous structure in Fig. 4c is quite open because only 43% of the space is occupied by silicon dioxide molecules. The relatively open structure accounts for the lower density and allows a variety of impurities (such as sodium) to enter and diffuse readily through the silicon dioxide layer.

The kinetics of thermal oxidation of silicon can be studied using the simple model illustrated² in Fig. 5. A silicon slice contacts the oxidizing species (oxygen or water vapor), resulting in a surface concentration of C_0 molecules/cm³ for these species. The magnitude of C_0 equals the equilibrium bulk concentration of the species at the oxidation temperature. The equilibrium concentration generally is proportional to the partial pressure of the oxidant adjacent to the oxide surface. At 1000 °C and a pressure of 1 atm, the concentration C_0 is 5.2×10^{16} molecules/cm³ for dry oxygen and 3×10^{19} molecules/cm³ for water vapor.

The oxidizing species diffuses through the silicon dioxide layer, resulting in a concentration C_s at the surface of silicon. The flux F_1 can be written as

$$F_1 = D \frac{dC}{dx} \cong \frac{D(C_0 - C_s)}{x}, \quad (3)$$

where D is the diffusion coefficient of the oxidizing species and x is the thickness of the oxide layer already present.

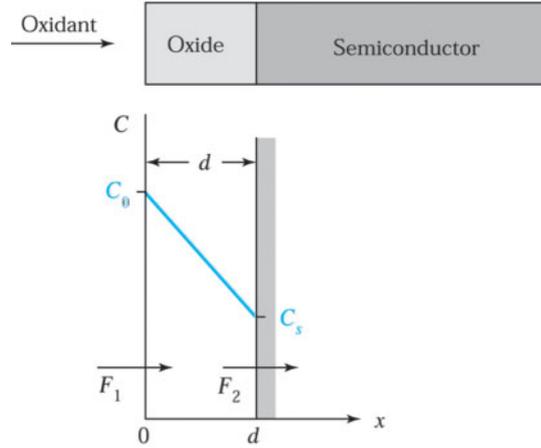


Fig. 5 Basic model for the thermal oxidation of silicon.²

At the silicon surface, the oxidizing species reacts chemically with silicon. Assuming the rate of reaction is proportional to the concentration of the species at the silicon surface, the flux F_2 is given by

$$F_2 = \kappa C_s, \quad (4)$$

where κ is the surface reaction rate constant for oxidation. At the steady state, $F_1 = F_2 = F$. Combining Eqs. 3 and 4 gives

$$F = \frac{DC_0}{x + (D/\kappa)}. \quad (5)$$

The reaction of the oxidizing species with silicon forms silicon dioxide. Let C_1 be the number of molecules of the oxidizing species in a unit volume of the oxide. There are 2.2×10^{22} silicon dioxide molecules/cm³ in the oxide, and we add one oxygen molecule (O₂) to each silicon dioxide molecule, whereas we add two water molecules (H₂O) to each silicon dioxide molecule. Therefore, C_1 for oxidation in dry oxygen is 2.2×10^{22} cm⁻³, and for oxidation in water vapor it is twice this number (4.4×10^{22} cm⁻³). Thus, the growth rate of the oxide layer thickness is given by

$$\frac{dx}{dt} = \frac{F}{C_1} = \frac{DC_0/C_1}{x + (D/\kappa)}. \quad (6)$$

We can solve this differential equation subject to the initial condition $x(0) = d_0$, where d_0 is the initial oxide thickness; d_0 can also be regarded as the thickness of oxide layer grown in an earlier oxidation step. Solving Eq. 6 yields the general relationship for the oxidation of silicon:

$$x^2 + \frac{2D}{\kappa}x = \frac{2DC_0}{C_1}(t + \tau), \quad (7)$$

where $\tau \equiv (d_0^2 + 2Dd_0/\kappa)C_1/2DC_0$, which represents a time coordinate shift to take into account the initial oxide layer d_0 .

The oxide thickness after an oxidizing time t is given by

$$x = \frac{D}{\kappa} \left[\sqrt{1 + \frac{2C_0\kappa^2(t + \tau)}{DC_1}} - 1 \right]. \quad (8)$$

For small values of t , Eq. 8 reduces to

$$x \cong \frac{C_0 \kappa}{C_1} (t + \tau), \quad (9)$$

and for larger values of t , it reduces to

$$x \cong \sqrt{\frac{2DC_0}{C_1}} (t + \tau). \quad (10)$$

During the early stages of oxide growth, when the surface reaction is the rate-limiting factor, the oxide thickness varies linearly with time. As the oxide layer becomes thicker, the oxidant must diffuse through the oxide layer to react at the silicon-silicon dioxide interface and the reaction becomes diffusion limited. The oxide growth then becomes proportional to the square root of the oxidizing time, which results in a parabolic growth rate.

Equation 7 is often written in a more compact form:

$$x^2 + Ax = B(t + \tau). \quad (11)$$

where $A = 2D/\kappa$, $B = 2DC_0/C_1$ and $B/A = \kappa C_0/C_1$. Using this form, Eqs. 9 and 10 can be written as

$$x = \frac{B}{A} (t + \tau) \quad (12)$$

for the linear region and as

$$x^2 = B(t + \tau). \quad (13)$$

for the parabolic region. For this reason, the term B/A is referred to as the linear rate constant and B as the parabolic rate constant. Experimentally measured results agree with the predictions of this model over a wide range of oxidation conditions. For wet oxidation, the initial oxide thickness d_0 is very small, or $\tau \cong 0$. However, for dry oxidation, the extrapolated value of d_0 at $t = 0$ is about 20 nm.

The temperature dependence of the linear rate constant B/A is shown in Fig. 6 for both dry and wet oxidation and for (111)- and (100)-oriented silicon wafers.² The linear rate constant varies as $\exp(-E_a/kT)$, where the activation energy E_a is about 2 eV for both dry and wet oxidation. This agrees closely with the energy required to break silicon-silicon bonds, 1.83 eV/molecule. Under a given oxidation condition, the linear rate constant depends on crystal orientation. This is because the rate constant is related to the rate of incorporation of oxygen atoms into the silicon. The rate depends on the surface bond structure of silicon atoms, making it orientation dependent. Because the density of available bonds on the (111)-plane is higher than that on the (100)-plane, the linear rate constant for (111)-silicon is larger.

Figure 7 shows the temperature dependence of the parabolic rate constant B , which can also be described by $\exp(-E_a/kT)$. The activation energy E_a is 1.24 eV for dry oxidation. The comparable activation energy for oxygen diffusion in fused silica is 1.18 eV. The corresponding value for wet oxidation, 0.71 eV, compares favorably with the value of 0.79 eV for the activation energy of diffusion of water in fused silica. The parabolic rate constant is independent of crystal orientation. This independence is expected because it is a measure of the diffusion process of the oxidizing species through a random network layer of amorphous silica.

Although oxides grown in dry oxygen have the best electrical properties, considerably more time is required to grow the same oxide thickness at a given temperature in dry oxygen than in water vapor. For relatively thin oxides such as the gate oxide in a MOSFET (typically ≤ 20 nm), dry oxidation is used.

However, for thicker oxides such as field oxides (≥ 20 nm) in MOS integrated circuits and for bipolar devices, oxidation in water vapor (or steam) is used to provide both adequate isolation and passivation.

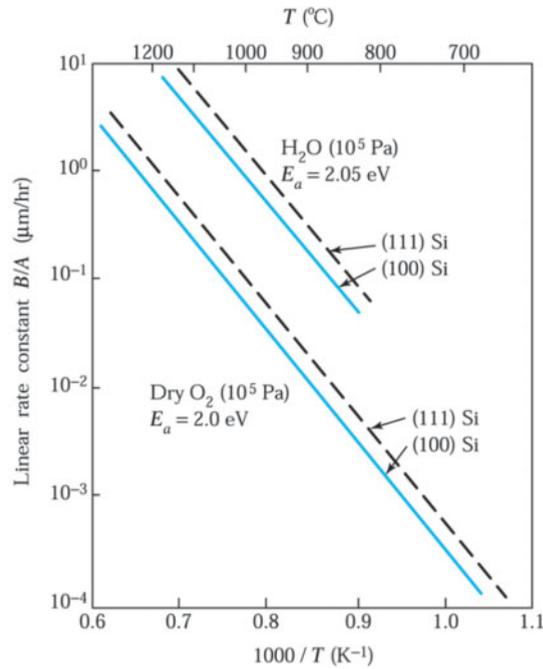


Fig. 6 Linear rate constant versus temperature.²

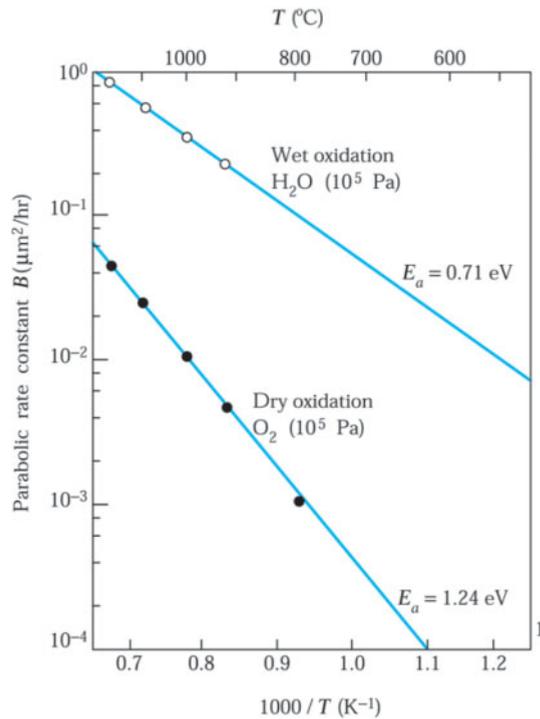


Fig. 7 Parabolic rate constant versus temperature.²

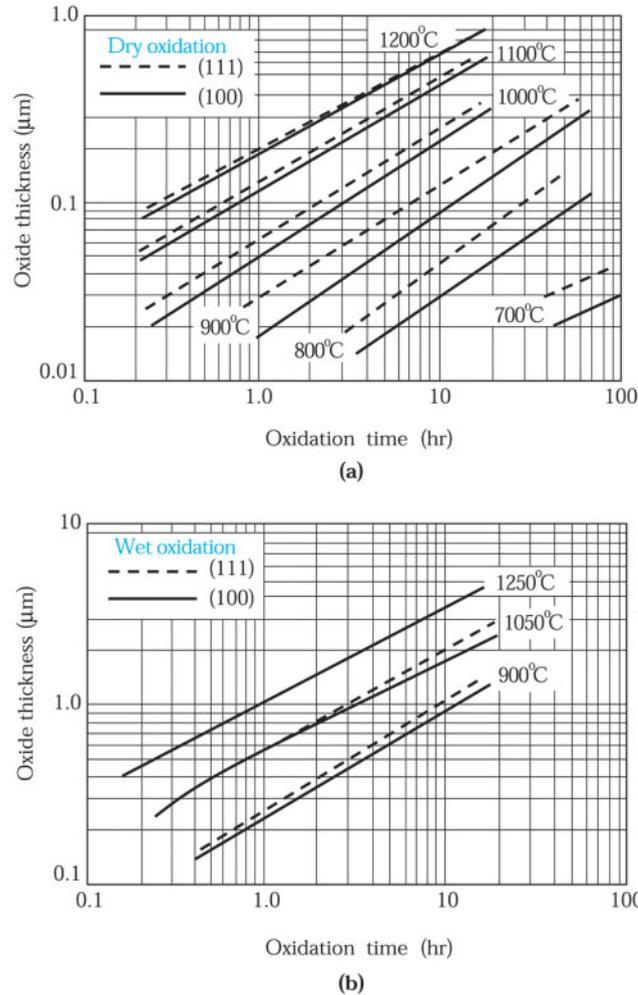


Fig. 8 Experimental results for silicon dioxide thickness as a function of reaction time and temperature for two substrate orientations. (a) Growth in dry oxygen. (b) Growth in steam.³

Figure 8 shows experimental results for silicon dioxide thickness as a function of reaction time and temperature for two substrate orientations.³ Under a given oxidation condition, the oxide thickness grown on a (111)-substrate is larger than that grown on a (100)-substrate because of the larger linear rate constant of the (111)-orientation. Note that for a give temperature and time, the oxide film obtained using wet oxidation is about 5–10 times thicker than that obtained using dry oxidation.

► EXAMPLE 2

Using Fig. 8, determine the thickness of an SiO_2 layer grown on a (100) bare Si wafer in the following three sequential steps: (a) 60 min., 1200 °C, dry O_2 , (b) 18 min., 900 °C, stream, (c) 30 min., 1050 °C, stream.

SOLUTION

(a) Since we are beginning with a bare silicon wafer, we can use Fig. 8a directly. We find a value of 0.18 μm or 180 nm.

- (b) Using $0.18\ \mu\text{m}$ as a starting point on Fig. 8b, we find that we have grown the equivalent of 0.7 hr or 42 min. We add another 18 min, bringing the total time to 60 min. Figure 8b shows a total oxide thickness of $0.22\ \mu\text{m}$.
- (c) Using $0.22\ \mu\text{m}$ as a starting point on Fig. 8b, we find that we have grown the equivalent of 15 min. We add another 30 min, bringing the total time to 45 min. Figure 8b shows a total oxide thickness of $0.48\ \mu\text{m}$.

12.1.2 Thin Oxide Growth

Relatively slow growth rates must be used to grow thin oxide films of precise thicknesses reproducibly. Various approaches to achieving such slower growth rates have been reported. The mainstream approach for gate oxides 10–15 nm thick is to grow the oxide film at atmospheric pressure and lower temperatures ($800^{\circ}\text{--}900^{\circ}\text{C}$). With this approach, processing using modern *vertical* oxidation furnaces can grow reproducible, high-quality 10 nm oxides to within 0.1 nm across the wafer.

We noted earlier that for dry oxidation, there is an apparently rapid oxidation that gives rise to an initial oxide thickness d_0 of about 20 nm. Therefore, the simple model presented in Section 12.1.1 is not valid for dry oxidation with oxide thickness ≤ 20 nm. For ultralarge-scale integration (ULSI), the ability to grow thin (5–20 nm), uniform, high-quality reproducible gate oxides has become increasingly important. We briefly consider the growth mechanisms of such thin oxides.

In the early stage of growth in dry oxidation, there is a large compressive stress in the oxide layer that reduces the oxygen diffusion coefficient in the oxide. As the oxide becomes thicker, the stress will be reduced due to the viscous flow of silica and the diffusion coefficient will approach its stress-free value. Therefore, for thin oxides, the value of D/κ may be sufficiently small that we can neglect the term Ax in Eq. 11 and obtain

$$x^2 - d_0^2 = Bt, \quad (14)$$

where d_0 is equal to $\sqrt{2DC_0\tau/C_1}$, which is the initial oxide thickness when time is extrapolated to zero, and B is the parabolic rate constant defined previously. We therefore expect the initial growth in dry oxidation to follow a parabolic form.

▶ 12.2 CHEMICAL VAPOR DEPOSITION OF DIELECTRICS

Deposited dielectric films are used mainly for insulation and passivation of discrete devices and integrated circuits. Considerations in selecting a deposition process are the substrate temperature, the deposition rate and film uniformity, the morphology, the electrical and mechanical properties, and the chemical composition of the dielectric films.

12.2.1 Chemical Vapor Deposition

Chemical vapor deposition (CVD) is the most useful method for the deposition of a wide variety of thin films in semiconductor device fabrication. CVD is used to deposit, for example, polysilicon for gate conductor, silica glass, doped silica glass such as borophosphosilicate glass (BPSG) and phosphosilicate glass (PSG), silicon nitride for dielectric films, and tungsten, tungsten silicide, and titanium nitride for conducting films. Other emerging dielectrics such as high-dielectric-constant materials (e.g., hafnium silicate), low-dielectric-constant materials (e.g., carbon-doped silicate glass), and conductors (e.g., copper barrier/tantalum nitride, copper, ruthenium) can also be deposited by CVD.

There are three commonly used deposition methods: atmospheric-pressure CVD, low-pressure CVD (LPCVD), and plasma-enhanced chemical vapor deposition (PECVD, or plasma deposition). The reactor for atmospheric-pressure CVD is similar to the one shown in Fig. 2, except that different gases are used at the gas inlet. LPCVD is a CVD process operated at subatmospheric pressures. Reduced pressures can reduce unwanted gas-phase reactions and improve film uniformity across the wafer. However, it suffers from low deposition

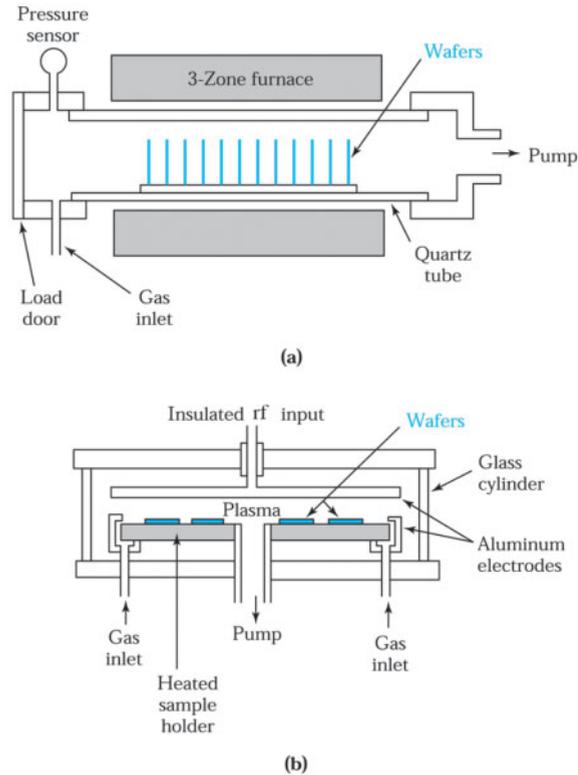


Fig. 9 Schematic diagrams of chemical-vapor deposition reactors. (a) Hot-wall LPCVD reactor. (b) Parallel-plate rf (radio frequency) plasma deposition reactor.⁴

rates. In a hot-wall LPCVD reactor as shown in Fig. 9a, the quartz tube is heated by a three-zone furnace, and gas is introduced at one end and pumped out at the opposite end. The semiconductor wafers are held vertically in a slotted quartz boat.⁴ The quartz tube wall is hot because it is adjacent to the furnace, in contrast to a cold-wall reactor such as the horizontal epitaxial reactor, which uses radio frequency (rf) heating. The choice of a hot-wall or cold-wall reactor depends on whether the reaction is exothermic or endothermic. For the exothermic reaction, the deposition rate is lower with increasing temperature. These processes require a hot-wall reactor. However, in a cold-wall reactor, deposition would occur on the cooler reactor walls. Consequently, for the endothermic reaction, a cold-wall reactor is used. The deposition rate is higher on the substrates with higher temperatures.

PECVD is an energy-enhanced CVD method in which plasma energy is added to the thermal energy of a conventional CVD system. The parallel-plate, radial-flow PECVD reactor shown in Fig. 9b consists of a cylindrical glass or aluminum chamber sealed with aluminum endplates. Inside are two parallel aluminum electrodes. An rf voltage is applied to the upper electrode, whereas the lower electrode is grounded. The rf voltage causes a plasma discharge between the electrodes. Wafers are placed on the lower electrode, which is heated to between 100° and 400°C by resistance heaters. The reaction gases flow through the discharge from inlets located along the circumference of the lower electrode. The main advantage of this reactor is its low deposition temperature. However, its capacity is limited, especially for large-diameter wafers, and the wafers may become contaminated if loosely adhering deposits fall on them.

The substrate surface not only receives active precursors but is subject to the bombardment of charged species. The short-lived active species react and deposit on the surface, while the thermal energy and ion bombardment continue to modify the deposited materials. The plasma-enhanced deposited films tend to be of smaller grain size or even amorphous, and contain amounts of impurities such as hydrogen, carbon or halide atoms.

The combination of low temperature, self-cleaning capability, and versatile film tunability has assured the importance of PECVD in the semiconductor industry. To minimize deposits on the reactor surfaces, limiting the plasma area is beneficial. The standard parallel-plate configuration provides an efficient design to focus the deposition on the wafer. At the same time, the reactor's plasma capability also provides the potential for in-situ plasma cleaning by introducing etchant cleaning gases such as C_2F_6 or NF_3 to remove silicon dioxide and silicon nitride deposition from chamber surfaces. One limitation of plasma deposition involves the potential charge imbedded in the film.

To overcome charge damage and still maintain the advantages of a low-temperature process, remote plasma instead of in-situ plasma is used. Reactants are plasma dissociated or activated remotely, then introduced onto the substrate surface along with second reactants to complete the reaction. But one has to consider the short lifetime of the activated species and how to distribute them over the large substrate surface. There is one closely related successful example, TEOS/ O_3 . Fortunately, the O_3 is stable enough and the concentration can be high enough to produce a reasonable silica deposition rate and provide good step coverage.

CVD Processes

Chemical vapor deposition (CVD) is a method of forming a thin solid film on a substrate by the reaction of vapor-phase chemicals that contain the required constituents. The CVD process can be generalized in a sequence of steps. (1) Reactants are introduced into the reactor; (2) Gas species are activated and dissociated by mixing, heating, plasma, or other means; (3) Reactive species are adsorbed on the substrate surface; (4) Adsorbed species undergo chemical reaction or react with other incoming species to form a solid film; (5) Reaction byproducts are desorbed from the substrate surface; (6) Reaction byproducts are removed from the reactor.

Although film growth is primarily accomplished at step 4, the overall growth rate is controlled by steps 1-6 in series. The slowest step determines the final growth rate. As in any typical chemical kinetics, the determining factors are the concentrations of surface species, wafer temperature, and incoming charged species and their energies. Chemical vapor deposition process parameters must be finely adjusted to meet all the film properties and production requirements.

12.2.2 Silicon Dioxide

CVD silicon dioxide does not replace thermally grown oxides because the best electrical properties are obtained with thermally grown films. CVD oxides are used instead to complement thermal oxides. A layer of undoped silicon dioxide is used to insulate multilevel metallization, to mask ion implantation and diffusion, and to increase the thickness of thermally grown field oxides. Phosphorus-doped silicon dioxide is used both as an insulator between metal layers and as a final passivation layer over devices. Oxides doped with phosphorus, arsenic, or boron are used occasionally as diffusion sources.

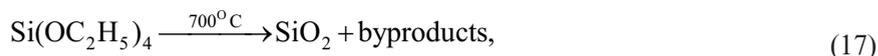
Deposition Methods

Silicon dioxide films can be deposited by several methods. For low-temperature deposition ($300^\circ\text{--}500^\circ\text{C}$), the films are formed by reacting silane (SiH_4), dopant, and oxygen. The chemical reactions for phosphorus-doped oxides are



The deposition process can be performed either at atmospheric pressure CVD or at LPCVD (Fig. 9a). The low deposition temperature of the silane-oxygen reaction makes it a suitable process when films must be deposited over a layer of aluminum.

For intermediate-temperature deposition ($500^\circ\text{--}800^\circ\text{C}$), silicon dioxide can be formed by decomposing tetraethylorthosilicate, $Si(OC_2H_5)_4$, in an LPCVD reactor. The compound, abbreviated TEOS, is vaporized from a liquid source. The TEOS compound decomposes as follows:

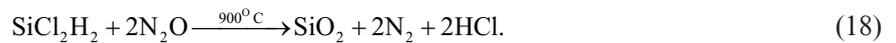


forming both SiO₂ and a mixture of organic and organosilicon byproducts. Although the higher temperature required for the reaction prevents its use over aluminum, it is suitable for polysilicon gates requiring a uniform insulating layer with good step coverage. The good step coverage is a result of enhanced surface mobility at higher temperatures. The oxides can be doped by adding small amounts of the dopant hydrides (phosphines, arsine, or diborane), similar to the process in epitaxial growth.

The deposition rate as a function of temperature varies as $e^{-E_a/kT}$, where E_a is the activation energy. The E_a of the silane-oxygen reaction is quite low: about 0.6 eV for undoped oxides and almost zero for phosphorus-doped oxide. In contrast, E_a for the TEOS reaction is much higher: about 1.9 eV for undoped oxide and 1.4 eV when phosphorus doping compounds are present. The dependence of the deposition rate on TEOS partial pressure is proportional to $(1 - e^{-P/P_0})$, where P is the TEOS partial pressure and P_0 is about 30 Pa. At low TEOS partial pressures, the deposition rate is determined by the rate of the surface reaction. At high partial pressures, the surface becomes nearly saturated with adsorbed TEOS and the deposition rate becomes essentially independent of TEOS pressure.⁴

Recently, atmospheric-pressure and low-temperature CVD processes using TEOS and ozone (O₃) have been proposed.⁵ This CVD technology produces oxide films with high conformality and low viscosity at low deposition temperatures. Because of their porosity, TEOS/O₃ CVD oxides are often accompanied by plasma-assisted oxides to permit planarization in ULSI processing.

For high-temperature deposition (900°C), silicon dioxide is formed by reacting dichlorosilane, SiCl₂H₂, with nitrous oxide at reduced pressure:



This deposition gives excellent film uniformity and is sometimes used to deposit insulating layers over polysilicon.

Properties of Silicon Dioxide

Deposition methods and properties of silicon dioxide films are listed⁴ in Table 1. In general, there is a direct correlation between deposition temperature and film quality. At higher temperatures, deposited oxide films are structurally similar to silicon dioxide that has been thermally grown.

The lower densities occur in films deposited below 500°C. Heating deposited silicon dioxide at temperatures between 600° and 1000°C causes densification, during which the oxide thickness decreases whereas the density increases to 2.2 g/cm³. The refractive index of silicon dioxide is 1.46 at a wavelength of 0.6328 μm. Oxides with lower indices are porous, such as the oxide from the silane-oxygen deposition, which has a refractive index of 1.44. The porous nature of the oxide also is responsible for the lower dielectric strength and hence a higher leakage current in the oxide film. The etch rates of oxides in a hydrofluoric acid solution depend on deposition temperature, annealing history, and dopant concentration. Usually higher-quality oxides are etched at lower rates.

TABLE 1 PROPERTIES OF SiO₂ FILMS

| Property | Thermally grown at 1000°C | SiH ₄ + O ₂ at 450°C | TEOS at 700°C | SiCl ₂ H ₂ + N ₂ O at 900°C |
|--|---------------------------|--|------------------|--|
| Composition | SiO ₂ | SiO ₂ (H) | SiO ₂ | SiO ₂ (Cl) |
| Density(g/cm ³) | 2.2 | 2.1 | 2.2 | 2.2 |
| Refractive index | 1.46 | 1.44 | 1.46 | 1.46 |
| Dielectric strength (10 ⁶ V/cm) | >10 | 8 | 10 | 10 |
| Etch rate (Å/min) (100:1 H ₂ O:HF) | 30 | 60 | 30 | 30 |
| Etch rate (Å/min) (buffered HF) | 440 | 1200 | 450 | 450 |
| Step coverage | — | Nonconformal | Conformal | Conformal |

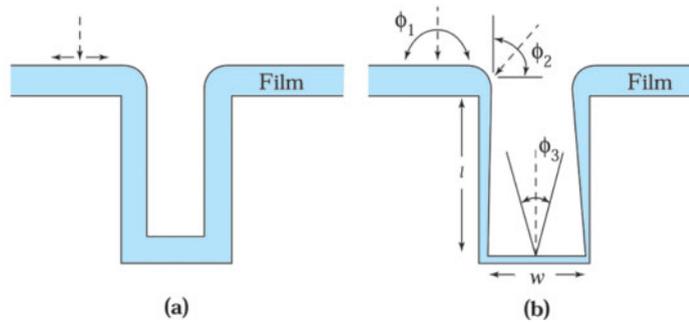


Fig. 10 Step coverage of deposited films. (a) Conformal step coverage. (b) Nonconformal step coverage.⁴

Step Coverage

Step coverage relates the surface topography of a deposited film to the various steps on the semiconductor substrate. Step coverage is one of the main advantages of the CVD method, especially compared with PVD. To get good step coverage, the inherent chemistries and operating conditions are critical. In the illustration of ideal, or conformal, step coverage shown in Fig. 10a, film thickness is uniform along all surfaces of the step. The uniformity of the film thickness, regardless of topography, is due to the rapid migration of reactants after adsorption on the step surfaces.⁶

Figure 10b shows an example of nonconformal step coverage, which results when the reactants adsorb and react without significant surface migration. In this instance, the deposition rate is proportional to the arrival angle of the gas molecules. Reactants arriving along the top horizontal surface come from many different angles and ϕ_1 , the arrival angle, varies in two dimensions from 0 to 180°, whereas reactants arriving at the top of a vertical wall have an arrival angle ϕ_2 that varies from 0° to 90°. Thus, the film thickness on the top surface is double that on a wall surface. Further down the wall, ϕ_3 is related to the width of the opening, and the film thickness is proportional to

$$\phi_3 \cong \arctan \frac{W}{l}, \quad (19)$$

where l is the distance from the top surface and W is the width of the opening. This type of step coverage is thin along the vertical walls, with a possible crack at the bottom of step caused by self-shadowing.

Silicon dioxide formed by TEOS decomposition at reduced pressure gives a nearly conformal coverage due to rapid surface migration. Similarly, the high-temperature dichlorosilane-nitrous oxide reaction also results in conformal coverage. However, during silane-oxygen deposition, no surface migration takes place and the step coverage is determined by the arrival angle. Most evaporated or sputtered materials have a step coverage similar to that in Fig. 10b.

P-Glass Flow

A smooth topography is usually required for the deposited silicon dioxide used as an insulator between metal layers. If the oxide used to cover the lower metal layer is concave, circuit failure may result from an opening that may occur in the upper metal layer during deposition. Because phosphorus-doped silicon dioxide (P-glass) deposited at low temperatures becomes soft and flows upon heating, it provides a smooth surface and is often used to insulate adjacent metal layers. This process is called P-glass flow. In addition, the phosphorus can further getter sodium to prevent its penetration to sensitive gate areas.

Figure 11 shows four cross sections of scanning-electron-microscope photographs of P-glass covering a polysilicon step.⁶ All samples are heated in steam at 1100°C for 20 min. Figure 11a shows a sample of glass that contains a negligibly small amount of phosphorus and does not flow. Note the concavity of the film and that the corresponding angle θ is about 120°. Figures 11b, 11c, and 11d show samples of P-glass with progressively higher phosphorus contents up to 7.2 wt% (weight percent). In these samples the decreasing step angles of the P-glass layer indicate how flow increases with phosphorus concentration. P-glass flow depends on annealing time, temperature, phosphorus concentration, and the annealing ambient.⁶

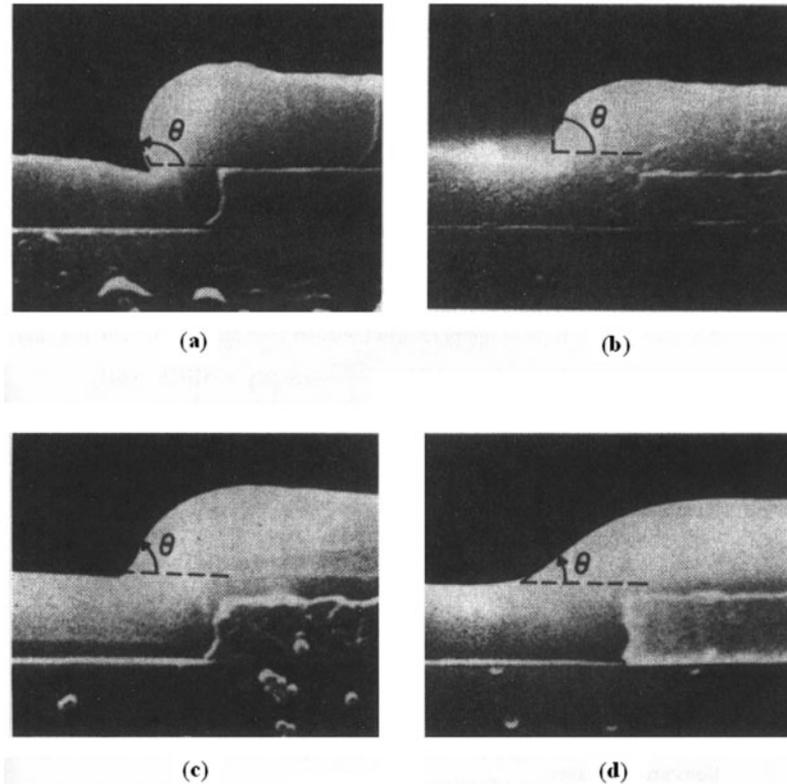


Fig. 11 Scanning-electron-microscope photographs (10,000 \times) of samples annealed in steam at 1100°C for 20 minutes for the following weight percent phosphorus: (a) 0 wt%; (b) 2.2 wt%; (c) 4.6 wt%; and (d) 7.2 wt%.

The angle θ as a function of weight percent of phosphorus as shown in Fig. 11 can be approximated by

$$\theta \cong 120^\circ \left(\frac{10 \text{ wt}\%}{10} \right). \quad (20)$$

If we want an angle smaller than 45° we require a phosphorus concentration larger than 6 wt%. However, at concentrations above 8 wt%, the metal film (e.g., aluminum) may be corroded by the acid products formed during the reaction between the phosphorus in the oxide and atmospheric moisture. Therefore, the P-glass flow process uses phosphorus concentrations of 6–8 wt%.

The efficiency of dopant incorporation is controlled by the decomposition mechanism of the dopant sources. In the thermal process, temperature is the dominant factor. In the plasma-enhanced process, the temperature dependence is much less, and plasma power is much more critical.

12.2.3 Silicon Nitride

It is difficult to grow silicon nitride by thermal nitridation (e.g., with ammonia, NH_3) because of its low growth rate and high growth temperature. However, silicon nitride films can be deposited by an intermediate-temperature (750°C) LPCVD process or a low-temperature (300°C) plasma-assisted CVD process.^{7,8} The LPCVD films are of stoichiometric composition (Si_3N_4) with high density (2.9–3.1 g/cm³). These films can be used to passivate devices because they serve as good barriers to the diffusion of water and sodium. The films also can be used as masks for the selective oxidation of silicon because silicon nitride oxidizes very slowly and prevents the underlying silicon from oxidizing. The films deposited by plasma-assisted CVD are not stoichiometric and have a lower density (2.4–2.8 g/cm³). Because of the low

deposition temperature, silicon nitride films can be deposited over fabricated devices and serve as their final passivation. The plasma-deposited nitride provides excellent scratch protection, serves as a moisture barrier, and prevents sodium diffusion.

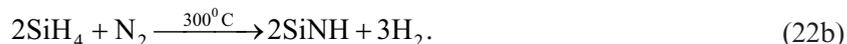
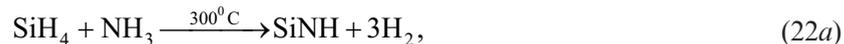
In the LPCVD process, dichlorosilane and ammonia react at reduced pressure to deposit silicon nitride at temperatures between 700° and 800°C. The reaction is



Good film uniformity and high wafer throughput (the number of wafers processed per hour) are advantages of the reduced-pressure process. As in oxide deposition, silicon nitride deposition is controlled by temperature, pressure, and reactant concentration. The activation energy for deposition is about 1.8 eV. The deposition rate increases with increasing total pressure or dichlorosilane partial pressure and decreases with increasing ammonia-to-dichlorosilane ratio.

Silicon nitride deposited by LPCVD is an amorphous dielectric containing up to 8 atomic percent (at%) hydrogen. The etch rate in buffered HF is less than 1 nm/min. The film has a very high tensile stress of approximately 10^{10} dynes/cm², which is nearly 10 times that of TEOS-deposited SiO₂. Films thicker than 200 nm may crack because of the very high stress. The resistivity of silicon nitride at room temperature is about 10^{16} Ω-cm. Its dielectric constant is 7 and its dielectric strength is 10^7 V/cm.

In the plasma-assisted CVD process, silicon nitride is formed either by reacting silane and ammonia in an argon plasma or by reacting silane in a nitrogen discharge. The plasma dissociates the precursors and creates high-energy forms of the reactant species that accelerate the reaction rate at a much lower temperature. Ions and electrons are charged species associated with plasma. The reactions are as follows:



The products depend strongly on deposition conditions. The radial-flow parallel-plate reactor (Fig. 9b) is used to deposit the films. The deposition rate generally increases with increasing temperature, power input, and reactant gas pressure.

Large concentrations of hydrogen are contained in plasma-deposited films. The plasma nitride (also referred to as SiN) used in semiconductor processing generally contains 20-25 at% hydrogen. Films with low tensile stress ($\sim 2 \times 10^9$ dynes/cm²) can be prepared by plasma deposition. Film resistivities range from 10^5 to 10^{21} Ω-cm, depending on silicon-to-nitrogen ratio, whereas dielectric strengths are between 1×10^6 and 6×10^6 V/cm. For passivation, the films must be a moisture and sodium diffusion barrier with good step coverage and no pinholes. Silicon nitride is an ideal material for a passivation layer, but high-temperature thermally deposited nitride exceeds the temperature for Al metallization and the hydrogen content in lower temperature PECVD nitride can cause a degradation in hot carrier lifetime.

12.2.4 Low-Dielectric-Constant Materials

As devices continue to scale down to the deep submicron region, they require multilevel interconnection architecture to minimize the time delay due to parasitic resistance R and capacitance C . The gain in device speed at the gate level will be offset by the propagation delay at the metal interconnects because of the increased RC time constant, as shown in Fig. 12. For example, in devices with gate length of 250 nm or less, up to 50% of the time delay is due to the RC delay of long interconnections.⁹ Therefore, the device interconnection network becomes a limiting factor in determining chip performance by affecting device speed, cross talk, and power consumption of ULSI circuits.

To reduce the RC time constant of ULSI circuits, interconnection materials with low resistivity and interlayer films with low capacitance are required. For low-capacitance ($C = \epsilon_i A/d$, where ϵ_i is the dielectric permittivity, A the area, and d the thickness of the dielectric film), it is not easy to lower the parasitic capacitance by increasing the thickness d of the interlayer dielectric (which makes gap filling more difficult), or decreasing wiring height and area A (which results in the increase of interconnect resistance). Therefore, materials with a low dielectric constant (low k) are required. The ϵ_i is equal to the product of k and ϵ_0 , where k and ϵ_0 are the dielectric constant and the vacuum permittivity, respectively.

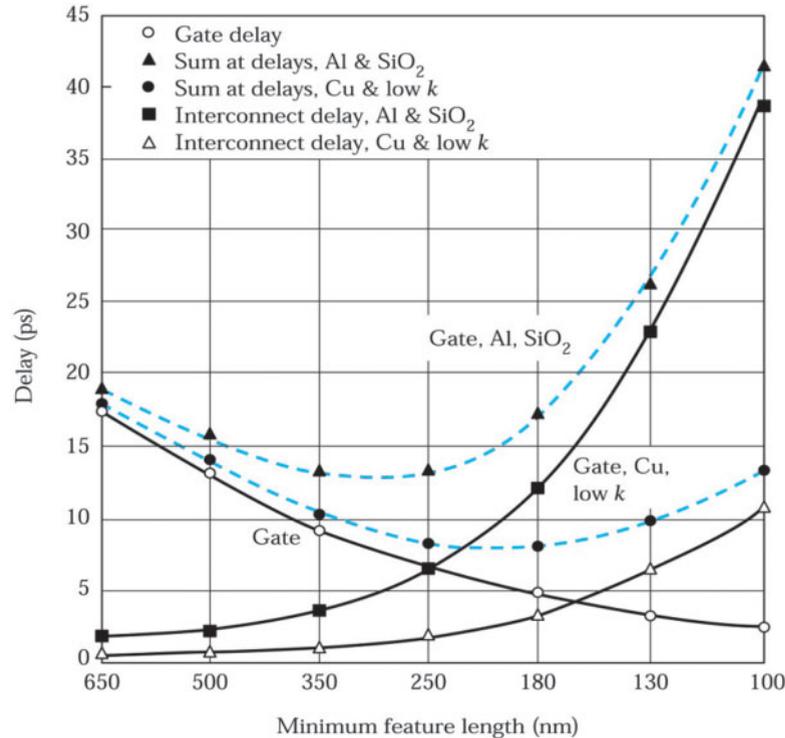


Fig. 12 Calculated gate and interconnect delay versus generation technology. The dielectric constant for the low-*k* material is 2.0. Both Al and Cu interconnects are 0.8 μm thick and 43 μm long.

Material Options

The properties of the interlayer dielectric film and how they are formed have to meet the following requirements: low dielectric constant, low residual stress, high planarization capability, high capability for gap filling, low deposition temperature, process simplicity, and ease of integration.

A substantial number of low-*k* materials have been synthesized for the intermetal dielectric in ULSI circuits. Some of the promising low-*k* materials are shown in Table 2. These materials can be either inorganic or organic and can be deposited by either CVD or spin-on techniques.⁹ CVD technique offers process flexibility. Bulk film and interface film properties can be easily altered in CVD processes by the adjustment of the process gas flow ratio or other process parameters, while those prepared by a spin-on technique can be changed only by modifying the precursor chemistry.

Basically, low-*k* materials are Si- and C-based, with very different characteristics. C-based materials (e.g., PAE, SiLK) generally have lower *k* values. Si-based materials (e.g., FSG, black diamond, HSQ, Xerogel) usually have higher thermal stability and hardness than C-based materials, but Si-based materials tend to be more prone to moisture absorption. Si-based materials are much more compatible with integration issues: adhesion to dielectrics and metals is better and they are easily etched with F-based etching chemistry and are more compatible with CMP processing.

Fluorine is one of the most electronegative elements. F in the silicate network would tie up electron density around itself, making the overall film less polarizable and hence reducing the dielectric constant.

There appear to be two possible migration paths for the future. The first is to continue with Si-based materials and introduce additional porosity into the film to reduce *k*. Possible disadvantages include lower mechanical strength and moisture absorption due to the porosity. The second path is to switch to C-based organic materials, which generally have lower *k* than Si-based materials. Which path will prevail depends on whether Si-based materials can prove extensibility to $k < 2.0$ or if the integration difficulties of C-based materials cannot be resolved in cost-effective ways.

TABLE 2 LOW-*k* MATERIALS

| Determinant | Materials | Dielectric constant |
|---------------------------------|-------------------------------------|---------------------|
| Vapor-phase deposition polymers | Fluorosilicate glass (FSG) | 3.5–4.0 |
| | Parylene N | 2.6 |
| | Parylene F | 2.4–2.5 |
| | Black diamond (C-doped oxide) | 2.7–3.0 |
| | Fluorinated hydrocarbon | 2.0–2.4 |
| | Teflon-AF | 1.93 |
| Spin-on polymers | HSQ/MSQ | 2.8–3.0 |
| | Polyimide | 2.7–2.9 |
| | SiLK (aromatic hydrocarbon polymer) | 2.7 |
| | Benzocyclobutenes | 2.6–2.7 |
| | PAE [poly(arylene ethers)] | 2.6 |
| | Fluorinated polyimides | 2.5–2.9 |
| | Fluorinated amorphous carbon | 2.1 |
| | Xerogels (porous silica) | 1.1–2.0 |

▶ EXAMPLE 3

Estimate the intrinsic RC value of two parallel Al wires $0.5 \mu\text{m} \times 0.5 \mu\text{m}$ in cross section, 1 mm in length, and separated by a polyimide ($k \sim 2.7$) dielectric layer that is $0.5 \mu\text{m}$ thick. The resistivity of Al is $2.7 \mu\Omega\text{-cm}$.

SOLUTION

$$RC = \left(\rho \frac{\ell}{t_m^2} \right) \times \left(\epsilon_i \frac{t_m \times \ell}{\text{spacing width}} \right) = \left(2.7 \times 10^{-6} \times \frac{1 \times 10^{-1}}{0.25 \times 10^{-8}} \right) \times \left(8.85 \times 10^{-14} \times 2.7 \times \frac{0.5 \times 10^{-4} \times 10^{-1}}{0.5 \times 10^{-4}} \right) \quad \blacktriangleleft$$

$$2.57 \text{ ps.}$$

12.2.5 High-Dielectric-Constant Materials

High- k materials are required for ULSI circuits, especially for dynamic random-access memory (DRAM). The storage capacitor in a DRAM has to maintain a certain value of capacitance for proper operation (e.g., 40 fF). For a given capacitance ($\epsilon_r A/d$), usually a minimum d is selected to meet the conditions of maximum allowed leakage current and minimum required breakdown voltage. The area of the capacitor can be increased by using stacked or trench structures. These structures are considered in Chapter 15. However, for a planar structure, area A is reduced with increasing DRAM density. Therefore, the dielectric constant of the film must be increased.

Several high- k materials have been proposed, such as barium strontium titanate (BST) and lead zirconium titanate (PZT). They are shown in Table 3. In addition, there are titanates doped with one or more acceptors, such as alkaline earth metals, or doped with one or more donors, such as rare earth elements. The tantalum oxide (Ta_2O_5) has a dielectric constant in a range of 20–30.

TABLE 3 HIGH- k MATERIALS

| | Materials | Dielectric constant |
|--------------------------|---|---------------------|
| Binary and quaternary | Ta ₂ O ₅ | 25 |
| | HfO ₂ | 18–22 |
| | HfSiON | 24 |
| | ZrO ₂ | 12–25 |
| | Al ₂ O ₃ | 9 |
| | TiO ₂ | 40–70 |
| | Y ₂ O ₃ | 17 |
| | Si ₃ N ₄ | 7 |
| Paraelectric perovskite | SrTiO ₃ (STO) | 140 |
| | (Ba _{1-x} Sr _x)TiO ₃ (BST) | 300–500 |
| | Ba(Ti _{1-x} Zr _x)O ₃ (BZT) | 300 |
| | (Pb _{1-x} La _x)(Zr _{1-y} Ti _y)O ₃ (PLZT) | 800–1000 |
| | Pb(Mg _{1/3} Nb _{2/3})O ₃ (PMN) | 1000–2000 |
| Ferroelectric perovskite | Pb(Zr _{0.47} Ti _{0.53})O ₃ (PZT) | >1000 |

As a reference, the dielectric constant of Si₃N₄ is in a range of 6–7 and that of SiO₂ is 3.9. A Ta₂O₅ film can be deposited by a CVD process using gaseous TaCl₅ and H₂O as the starting materials.

A Ta₂O₅ film can also be deposited by a thermal CVD process using metal-organic precursors, tantalum ethoxide (TAETO) or tantalum tetraethoxy dimethylaminoethoxide (TATDMAE), as the starting materials. For good step coverage the deposition process has to be carried out in the reaction-rate limited region. As-deposited TaO_x film is oxygen deficient and resistive in nature. Oxygen annealing of this film is essential for it to act as an effective dielectric material.

► EXAMPLE 4

A DRAM capacitor has the following parameters: capacitance $C = 40$ fF, cell size $A = 1.28 \mu\text{m}^2$, and dielectric constant $k = 3.9$ for silicon dioxide. If we replace SiO₂ with Ta₂O₅ ($k = 25$) without changing the thickness, what is the equivalent cell area of the capacitor?

SOLUTION

$$C = \frac{\epsilon_i A}{d},$$

$$\frac{3.9 \times 1.28}{d} = \frac{25 \times A}{d},$$

$$\therefore \text{Equivalent cell size } A = \frac{3.9}{25} \times 1.28 = 0.2 \mu\text{m}^2.$$

► 12.3 CHEMICAL VAPOR DEPOSITION OF POLYSILICON

Using polysilicon as the gate electrode in MOS devices is a significant development in MOS technology. One important reason is that polysilicon surpasses aluminum in electrode reliability. Figure 13 shows the maximum time to breakdown for capacitors with polysilicon and aluminum electrodes.¹⁰ The polysilicon is clearly superior,

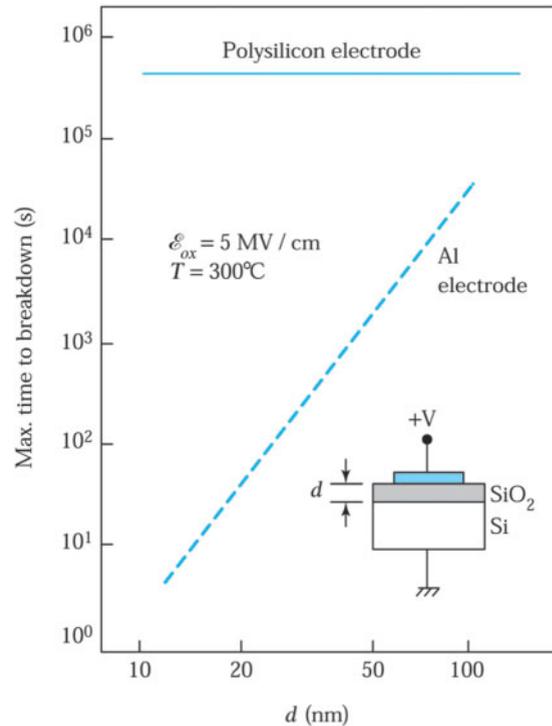


Fig. 13 Maximum time to breakdown versus oxide thickness for a polysilicon electrode and an aluminum electrode.¹⁰

especially for thinner gate oxides. The inferior time to breakdown of aluminum electrode is due to the migration of aluminum atoms into the thin oxide under an electrical field. Polysilicon is also used as a diffusion source to create shallow junctions and to ensure ohmic contact to crystalline silicon. Additional uses include the manufacture of conductors and high-value resistors.

A low-pressure reactor (Fig. 9a) operated between 600° and 650°C is used to deposit polysilicon by pyrolyzing silane according to the following reaction:



Of the two most common low-pressure processes, one operates at a pressure of 25–130 Pa using 100% silane, whereas the other process involves a diluted mixture of 20%–30% silane in nitrogen at the same total pressure. Both processes can deposit polysilicon on hundreds of wafers per run with good uniformity (i.e., thickness within 5%)

Figure 14 shows the deposition rate at four deposition temperatures. At low silane partial pressure, the deposition rate is proportional to the silane pressure.⁴ At higher silane concentrations, saturation of the deposition rate occurs. Deposition at reduced pressure is generally limited to temperatures between 600° and 650°C. In this temperature range, the deposition rate varies as $\exp(-E_a/kT)$, where the activation energy E_a is 1.7 eV, which is essentially independent of the total pressure in the reactor. At higher temperatures, gas-phase reactions that result in a rough, loosely adhering deposit become significant and silane depletion will occur, causing poor uniformity. At temperatures much lower than 600°C, the deposition rate is too slow to be practical.

Process parameters that affect the polysilicon structure are deposition temperature, dopants, and the heat cycle applied following the deposition step. A columnar structure results when polysilicon is deposited at temperatures of 600°–650°C. This structure is comprised of polycrystalline grains ranging in size from 0.03 to 0.3 μm at a preferred orientation of (110). When phosphorus is diffused at 950°C, the structure changes to crystallite and the grain size increases to between 0.5 and 1.0 μm . When the temperature is increased to 1050°C during oxidation, the grains reach a final size of 1–3 μm . Although the initially deposited film appears amorphous

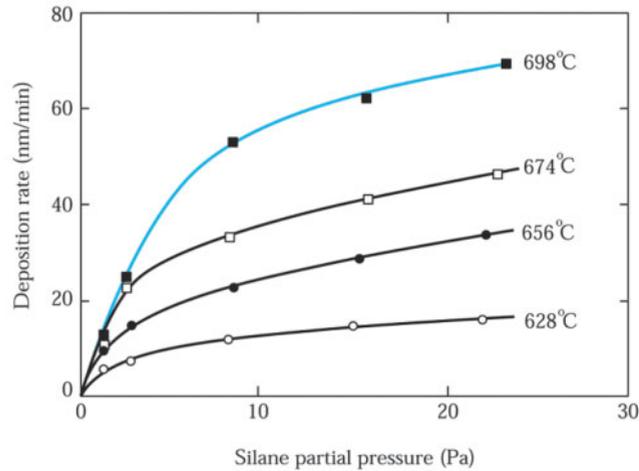


Fig. 14 Effect of silane concentration on the polysilicon deposition rate.⁴

when deposition occurs below 600°C, growth characteristics similar to the polycrystalline-grain columnar structure are observed after doping and heating.

Polysilicon can be doped by diffusion, ion implantation, or the addition of dopant gases during deposition, referred to as in-situ doping. The implantation method is most commonly used because of its lower processing temperatures. Figure 15 shows the sheet resistance of single crystal silicon and of 500 nm polysilicon doped with phosphorus and antimony using ion implantation.¹¹ The ion implantation process is considered in Chapter 14. Implant dose, annealing temperature, and annealing time all influence the sheet resistance of implanted polysilicon. Carrier traps at the grain boundaries cause a very high resistance in the lightly implanted polysilicon. As Fig. 15 illustrates, resistance drops rapidly, approaching that of implanted single crystal silicon, as the carrier traps become saturated with dopants.

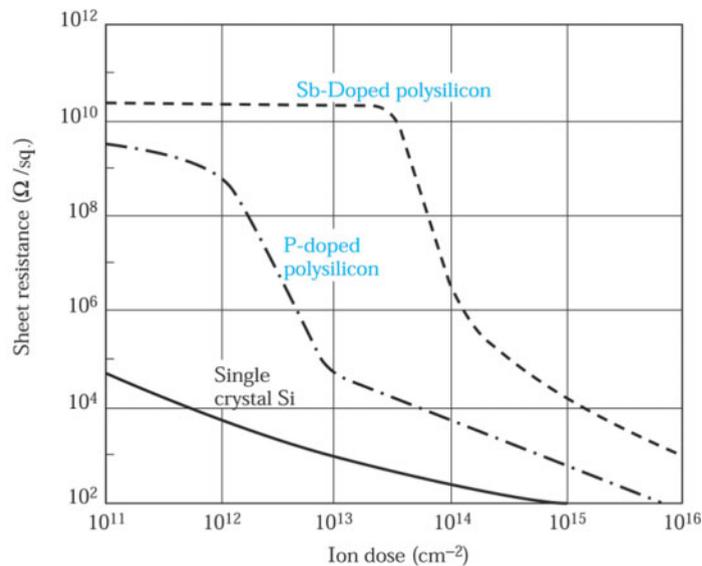


Fig. 15 Sheet resistance versus ion dose in 500 nm polysilicon at 30 keV.¹¹

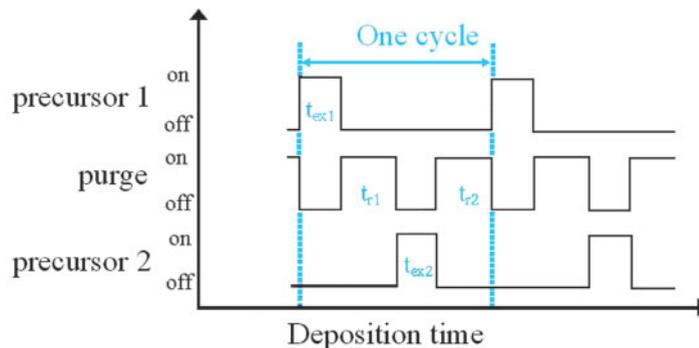


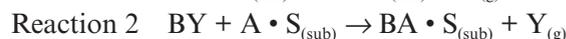
Fig. 16 A typical ALD cycle.

► 12.4 ATOMIC LAYER DEPOSITION

Atomic layer deposition (ALD) is a special chemical vapor deposition technique that is capable of depositing thin films of the order of a monolayer. ALD has emerged as an important method for nano-device fabrication, particularly for conformal coating on device structures with high aspect ratios from 20 - 100:1 at feature size below 100 nm.

ALD differs from conventional CVD in that CVD uses a continuous supply of chemical reactants that coexist in space and time above the semiconductor substrate. ALD uses sequential exposures of chemical reactants, each reactant having self-limiting deposition separated in time. In CVD, chemical reactions occur in the gas phase or on the substrate; but in ALD, chemical reactions take place only on the substrate and can prevent gas-phase reactions.

ALD is operated at low pressure. In an ALD binary thin film deposition, there are two sequential reactions.



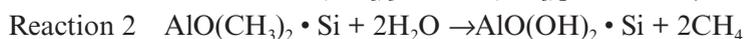
where AX is precursor 1, BY is precursor 2, $S_{(\text{sub})}$ is the substrate, and $X_{(\text{g})}$, $Y_{(\text{g})}$ are residuals.

A typical ALD cycle is shown in Fig. 16:

1. Expose precursor 1 for time (t_{ex1}) to carry out the first surface reaction.
2. Removal (purge) time (t_{r1}) of the unused precursor and reaction products of reaction 1.
3. Expose precursor 2 for time (t_{ex2}) to carry out the second surface reaction.
4. Removal (purge) time (t_{r2}) of the unused precursor and reaction products of reaction 2.

The cycle time can be as short as a fraction of a second or as long as a few minutes. The processes are repeated to build the film. The cycle time is defined as the sum of exposure and removal periods. Like CVD, ALD may be carried out by thermal reactions or by plasma-assisted processes.

We take the ALD- Al_2O_3 as an example to depict ALD growth processes. Figure 17 illustrates the two sequential reactions of ALD- Al_2O_3 using $\text{Al}(\text{CH}_3)_3$ (trimethylaluminum-TMA) as precursor 1 and H_2O as precursor 2.¹² Silicon is used as the substrate. The two sequential reactions of ALD- Al_2O_3 are:



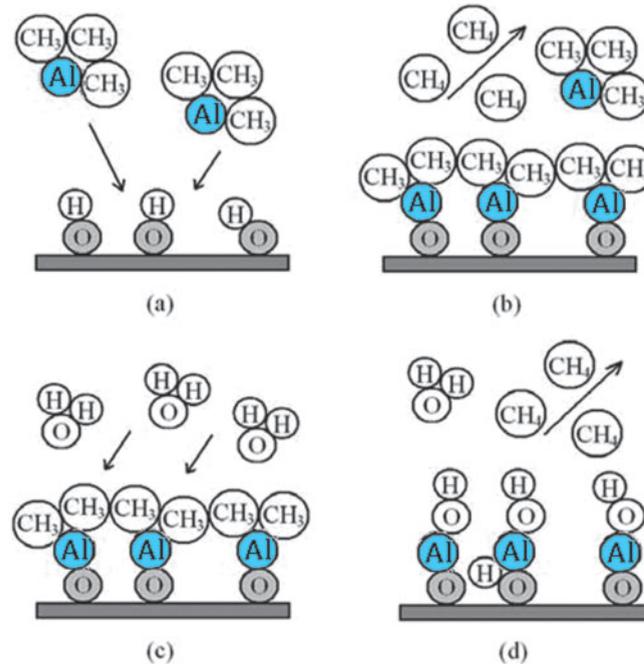


Fig. 17 (a) Reaction with OH (hydroxylated) surface exposed to TMA. (b) Removal of CH_4 byproducts and unused TMA reactant by chemical reaction. (c) Reaction with a CH_3 terminated surface exposed to H_2O . (d) Removal of CH_4 byproduct and unused H_2O reactant by chemical reaction.¹²

The reactions are repeated to build the ALD- Al_2O_3 film. The “ALD window” is the temperature range in which the deposition rate ($\text{\AA}/\text{cycle}$) is a constant, independent of the deposition temperature, as shown in Fig. 18. At lower temperatures, there is insufficient energy to achieve a complete chemical reaction. The chemical adsorption-reaction dominates and the deposition rate increases with temperature. At higher temperatures, a region of desorption dominates and the deposition rate decreases with temperature.

A non-ALD deposition associated with condensation phenomena is shown at the upper left at lower temperatures. Additionally, deposition by pyrolytic CVD from the decomposition of precursors at higher temperatures is shown in the upper right. The deposition rates of these processes may be higher than that of the ALD process.

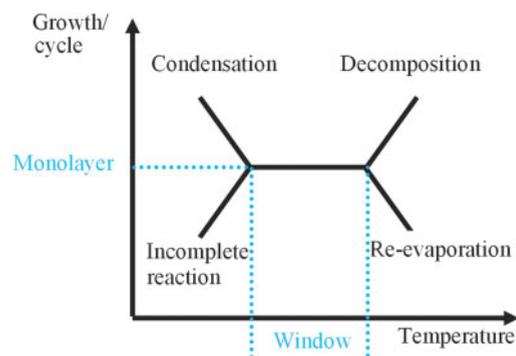


Fig. 18 The temperature dependence of the ALD deposition rate and related processes.

In ALD, film thickness depends only on the number of reaction cycles, which makes thickness control accurate and simple. There is less need of reactant flux homogeneity than in CVD. Therefore, ALD can give large area (large batch and easy scale-up) capability and excellent conformality and reproducibility. ALD can be used to deposit several types of thin films, including oxides (e.g., Al_2O_3 , TiO_2 , SnO_2 , ZnO , HfO_2), metal nitrides (e.g. TiN , TaN , WN , NbN), metals (e.g. Ru , Ir , Pt), and metal sulfides (e.g. ZnS). ALD has potentials in three mainstream applications: capacitors, gates, and interconnects. The major limitation of ALD is its low deposition rate; usually only a fraction of a monolayer is deposited in one cycle. Fortunately, the films needed for future-generation ICs are very thin and thus ALD's low deposition rate is not such an important issue.

► 12.5 METALLIZATION

12.5.1 Physical-Vapor Deposition

The primary semiconductor applications of physical-vapor deposition (PVD) technology are the deposition of metal and compounds such as Ti , Al , Cu , TiN , and TaN for lines, pads, vias, contacts, and related connections that are used to connect with the junctions and devices on the Si wafer surface.

The most common methods of PVD of metals are evaporation, e-beam evaporation, plasma spray deposition, and sputtering. Evaporation occurs when a source material is heated above its melting point in an evacuated chamber. The evaporated atoms then travel at high velocity in straight-line trajectories. The source can be made molten by resistance heating, by rf heating, or with a focused electron beam. Evaporation and e-beam evaporation were used extensively in earlier generations of integrated circuits, but they have been replaced for ULSI circuits by sputtering due to its volatility and high film quality.

Sputtering involves the transport of material from a target to a substrate. It is accomplished by the bombardment of the target surface with gas ions, typically Ar but occasionally other inert gas species (Ne , Kr) or reactive species such as oxygen or nitrogen. Particles of atomic dimension from the target are ejected as a result of momentum transfer between incident ions and the target, as shown in Fig. 19. The process is analogous to the action of a billiard ball hitting another billiard ball.

There are basically two kinds of sputtering systems, dc and rf sputtering. The dc (direct current) sputtering is usually used for metal film deposition. Figure 20a shows the standard sputtering system. There are two electrodes in the dc sputtering system. As a negative dc bias is applied directly on the cathode electrode of the metal target, the stray electrons accelerate and gain energy from the electric field to bombard Ar neutral atoms. If the bombarding electrons have sufficiently higher energy than the argon ionization energy (i.e., 15.7 eV), Ar is ionized and plasma is created. The positive argon ions in plasma are accelerated toward the metal target and sputter metal atoms off. The glow region of the plasma is a good conductor. At the start of Ar gas breakdown, the voltage between the two electrodes drops and hardly sustains a high field for the generation of plasma. The secondary electrons emitted from the metal target during sputtering sustain the plasma.

For semiconductor applications, a magnetron sputtering based on the variation of dc sputtering has higher efficiency. The cathode in magnetron sputtering differs from a conventional planar cathode in that there is a local magnetic field parallel to the cathode surface. The effect of the tangential magnetic field can move the emitted secondary electrons to the cathode surface. These electrons are trapped close to the cathode region and can lead to very high levels of gas ionization, which increases the ion density and hence, the sputter-deposition rate.

Directional Deposition

Contact holes with large aspect ratio are difficult to fill with material, mainly because scattering events cause the top opening of the hole to seal before appreciable material has deposited on its floor. The fundamental problem of putting atoms into a deep feature can be solved by enhanced directionality of the atoms as they are deposited. There are two ways to enhance sputtering directionality: long-throw sputtering and collimated sputtering.

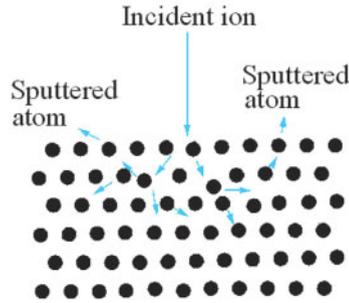


Fig. 19 Schematic of sputtering process.

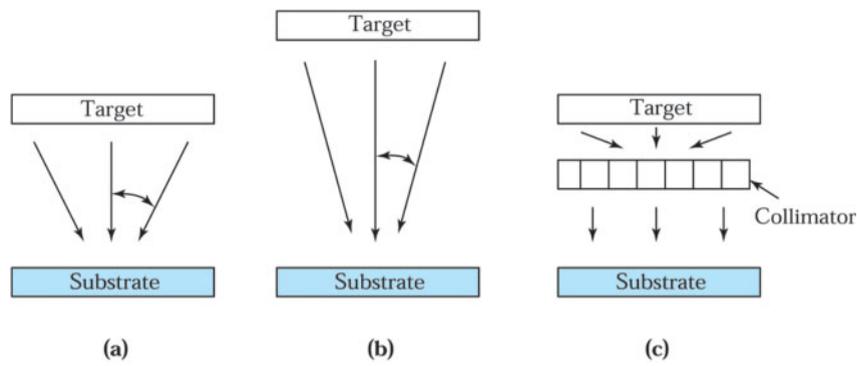


Fig. 20 (a) Standard sputtering, (b) long-throw sputtering, and (c) sputtering with a collimator.

Long-Throw Sputtering

By moving the sample farther away from the cathode for “long-throw” sputter deposition, as shown in Fig. 20b, an increasing fraction of the sputtered atoms is lost onto the sidewall of the chamber. This fraction is mainly determined by the target-to-substrate separation, d_{ts} , and scattering of the flux by the working gas. The larger d_{ts} is, the wider the angular distribution. The atoms arriving at the substrate are more likely to be closer to normal incidence than the conventional, short-throw deposition. The throw distance of the “long-throw” sputter deposition needs to be on the order of the cathode diameter. The process is limited in a practical sense by the gas scattering, which is associated with the operating pressure of the system. To reduce in-flight scattering, the mean free path for the sputtered atoms should exceed the throw distance. For “long-throw” sputtering deposition, the working pressure is very low (less than 0.1 Pa), again to reduce in-flight scattering. At such a low pressure, gas scattering is less important and the d_{ts} can be greatly increased. This allows more deposits at the bottom of high-aspect features such as contact holes.

Collimated Sputtering

In a long-mean-free-path deposition environment (mean free path > throw distance), geometric filtering of the deposition flux can be obtained by placing a collimator between the target and the sample. The collimator serves as a simple directional filter by collecting the atoms that impinge on its walls, as shown schematically in Figure 20c. The degree of filtering is simply a function of the aspect ratio of the collimator, where aspect ratio is defined as the thickness of the collimator divided by the diameter of a tube.

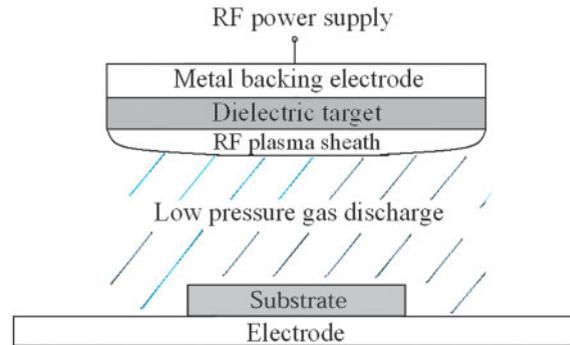


Fig. 21 Schematic diagram of RF sputtering.

RF sputtering

Rf (radio-frequency, typically 13.56 MHz, a frequency chosen because of its non-interference with radio-transmitted signals) sputtering is usually used in cases of dielectric materials, such as the high- k dielectrics. Figure 21 shows the standard rf sputtering system. It has several advantages: (a) its ability to sputter dielectrics as well as metals, (b) its ability to operate in the bias-sputtering mode, and (c) its ability to permit sputter-etching of substrates prior to deposition. When a time-varying potential is applied to a metal plate behind the dielectric target in rf sputtering, another time-varying potential is developed on the opposite target surface through the impedance of the target. Once the gas is broken down by the acceleration of stray electrons from the electric field to start a discharge, the current can flow from the plasma to the target surface. Since electrons are more mobile than positive ions, more electrons are attracted to the front surface of the target during the positive half cycle than are positive ions in the negative half cycle. Therefore, the current is larger in the positive cycle than that in the negative cycle, as in a diode. The resultant electron current causes the target surface to acquire an increasingly negative bias voltage during successive cycles until the negative average dc voltage is sufficiently high to retard the electrons' arrival, so that the net charge arriving at the target surface is zero.

Since the target potential is negative with respect to the plasma, electrons are forced away from the surface, yielding an ion sheath that is visible as a dark space (because there is no optical emission from the recombination of electrons and ions) near the target surface. Positive ions in the sheath are accelerated toward the target by the negative potential. To prevent accumulation of excessive positive ions at the target surface, the frequency of the applied voltage must be high. The frequency must be at least 10^6 Hz for any appreciable sputtering to occur. Below this frequency, the average energy of the ions is reduced significantly as a result of positive ions accumulating on the target.

RF-sputter etching is the reverse of the sputtering process, and is also known as back sputtering, reverse sputtering, ion etching, or sputter cleaning. The normal rf power flow is electrically reversed; the substrate has a negative average dc voltage and an anode takes the place of the target. RF-sputter etching is used to clean substrates prior to sputtering a film on them, or to make patterns on substrates.

Bias-sputtering is the bombardment by energetic positive ions of a growing film that has a negative bias. This technology can remove impurities on the growing film. Usually, it is used for substrate surface cleaning before dielectric film deposition.

12.5.2 CVD Metal Deposition

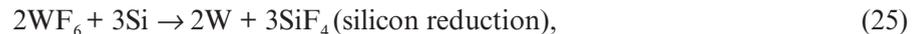
CVD is attractive for metallization because it offers coatings that are conformal, has good step coverage, and can coat a large number of wafers at a time. The basic CVD setup is the same as that used for the deposition of dielectrics and polysilicon (see Fig. 9a). Low-pressure CVD (LPCVD) is capable of producing conformal step coverage over a wider range of topographical profiles, often with lower electrical resistivity than that from PVD.

One of the major new applications of CVD metal deposition for integrated circuits is in the area of refractory-metal deposition. For example, tungsten's low electrical resistivity ($5.3 \mu\Omega\text{-cm}$) and its refractory nature make it a desirable metal for use in integrated circuits.

CVD Tungsten

Tungsten is used both as a contact plug and as a first-level metal. The CVD tungsten film is known for its excellent step coverage. For contact or via holes with size $< 0.8 \mu\text{m}$ and aspect ratios greater than two, it is difficult to use conventional Al sputtering for continuous coating inside the feature and maintain the electrical performance. The effective via resistance and electromigration resistance have been improved by the introduction of CVD tungsten. The CVD tungsten process has been a key technology enabling multilevel interconnection metallization.

Tungsten can be deposited by using WF_6 as the W source gas, since it is a liquid that boils at room temperature. WF_6 can be reduced by silicon, hydrogen, or silane. The basic chemistry for CVD-W is as follows:



On a Si contact, the selective process starts from a silicon reduction process. This process provides a nucleation layer of W grown on Si but not on SiO_2 . The hydrogen reduction process can deposit W rapidly on the nucleation layer, forming the plug. The hydrogen reduction process provides excellent conformal coverage of the topography. This process, however, does not have perfect selectivity, and the HF gas by-product of the reaction is responsible for the encroachment of the oxide, as well as for the rough surface of deposited W films.

The silane reduction process gives a high deposition rate and much smaller W grain size than that obtained with the hydrogen reduction process. In addition, the problems of encroachment and rough W surface are eliminated because there is no HF by-product. Usually, a silane reduction process is used as the first step in blanket W deposition to serve as a nucleation layer and to reduce junction damage. After the silane reduction process, hydrogen reduction is used to grow the blanket W layer.

CVD TiN

Titanium nitride (TiN) is widely used as a diffusion barrier-metal layer in metallization and has numerous applications: (1) a cladding layer in Al metallization to enhance interconnection wiring electromigration resistance, (2) a CVD-W adhesion layer over oxide and a barrier against the interaction of WF_6 with Al and Si, (3) local interconnection where Al metallization cannot bear the temperature, (4) a plate electrode for Ta_2O_5 capacitors, and (5) the node and plate electrodes for MIM (metal-insulator-metal) capacitors where the insulator is either an atomic-layer-deposited Al_2O_3 or $\text{HfO}_2/\text{Al}_2\text{O}_3$ laminate.

TiN can be deposited by sputtering from a compound target or by CVD. The CVD TiN can provide better step coverage than PVD methods in deep submicron technology. CVD TiN can be deposited,¹³⁻¹⁵ using TiCl_4 with NH_3 , H_2/N_2 , or NH_3/H_2 :



The deposition temperature is about $400^\circ\text{--}700^\circ\text{C}$ for NH_3 reduction and is above 700°C for the H_2/N_2 reaction. The higher the deposition temperature, the better the TiN film and the less Cl incorporated in the TiN ($\sim 5\%$).

12.5.3 Aluminum Metallization

Aluminum and its alloys are used extensively for metallization in integrated circuits. The Al film can be deposited by a PVD or CVD method. Because aluminum and its alloys have low resistivities ($2.7 \mu\Omega\text{-cm}$ for Al and up to $3.5 \mu\Omega\text{-cm}$ for its alloys), these metals satisfy the low-resistance requirements. Aluminum also adheres well to silicon dioxide. However, the use of aluminum in integrated circuits with shallow junctions often creates problems such as spiking and electromigration. We consider the problems of aluminum metallization and their solutions in this section.

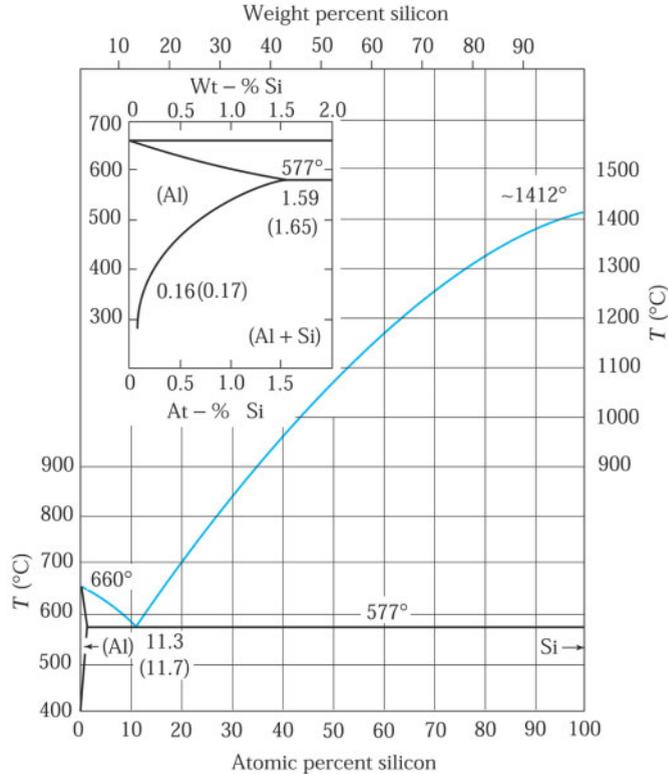


Fig. 22 Phase diagram of the aluminum-silicon system.¹⁶

Junction Spiking

Figure 22 shows the phase diagram of the Al-Si system at 1 atm.¹⁶ The phase diagram relates these two components as a function of temperature. The Al-Si system exhibits eutectic characteristics; that is, the addition of either component lowers the system's melting point below that of either metal. Here, the minimum melting temperature, called the eutectic temperature, is 577°C, corresponding to a 11.3% Si and 88.7% Al composition. The melting points of pure aluminum and pure silicon are 660°C and 1412°C, respectively. Because of the eutectic characteristics, during aluminum deposition the temperature on the silicon substrate must be limited to less than 577°C.

The inset of Fig. 22 also shows the solid solubility of silicon in aluminum. For example, the solubility of silicon in aluminum is 0.25 wt% at 400°C, 0.5 wt% at 450°C, and 0.8 wt% at 500°C. Therefore, wherever aluminum contacts silicon, the silicon will dissolve into the aluminum during annealing. The amount of silicon dissolved will depend not only on the solubility at the annealing temperature but also on the volume of aluminum to be saturated with silicon. Consider a long aluminum metal line in contact with an area ZL of silicon as shown in Fig. 23. After an annealing time t , the silicon will diffuse a distance of approximately \sqrt{Dt} along the aluminum line from the edge of the contact, where D is the diffusion coefficient given by $4 \times 10^{-2} \exp(-0.92/kT)$ for silicon diffusion in deposited aluminum films. Assuming that this length of aluminum is completely saturated with silicon, the volume of silicon consumed is then

$$\text{Vol} \cong 2\sqrt{Dt}(\text{HZ})S \left(\frac{\rho_{\text{Al}}}{\rho_{\text{Si}}} \right), \quad (30)$$

where ρ_{Al} and ρ_{Si} are the densities of aluminum and silicon, respectively, and S is the solubility of silicon in aluminum at the annealing temperature.¹⁷ If the consumption takes place uniformly over the contact area A (where $A = ZL$ for uniform dissolution), the depth to which silicon would be consumed is

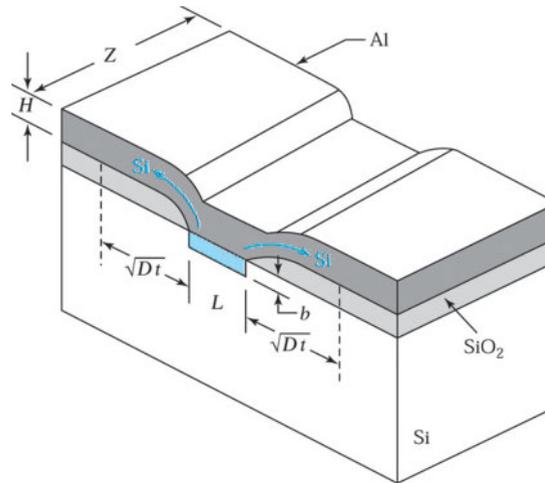


Fig. 23 Diffusion of silicon in aluminum metallization.¹⁷

$$b = 2\sqrt{Dt} \left(\frac{HZ}{A} \right) S \left(\frac{\rho_{Al}}{\rho_{Si}} \right). \quad (31)$$

► **EXAMPLE 5**

For $T = 500^\circ\text{C}$, $t = 30$ min, $ZL = 16 \mu\text{m}^2$, $Z = 5 \mu\text{m}$, and $H = 1 \mu\text{m}$. Find the depth b , assuming uniform dissolution.

SOLUTION The diffusion coefficient of silicon in aluminum at 500°C is about $2 \times 10^{-8} \text{ cm}^2/\text{s}$; thus, \sqrt{Dt} is $60 \mu\text{m}$. The density ratio is $2.7/2.33 = 1.16$. At 500°C , S is $0.8 \text{ wt}\%$. From Eq. 31 we have

$$b = 2 \times 60 \left(\frac{1 \times 5}{16} \right) 0.8\% \times 1.16 = 0.35 \mu\text{m}.$$

Aluminum will fill a depth of $b = 0.35 \mu\text{m}$ from which silicon is consumed. If at the contact point there is a shallow junction whose depth is less than b , the diffusion of silicon into aluminum can short-circuit the junction. ◀

In a practical situation, the dissolution of silicon does not take place uniformly but rather at only a few points. The effective area in Eq. 31 is less than the actual contact area; hence b is much larger. Figure 24 illustrates the actual situation in the p - n junction area of aluminum penetrating the silicon at only the few points where spikes are formed. One way to minimize aluminum spiking is to add silicon to the aluminum by coevaporation until the amount of silicon contained by the alloy satisfies the solubility requirement. Another method is to introduce a barrier metal layer between the aluminum and the silicon substrate (Fig. 25). This barrier metal layer must meet the following requirements: it forms low contact resistance with silicon, it will not react with aluminum, and its deposition and formation are compatible with the overall process. Barrier metals such as titanium nitride (TiN) have been evaluated and found to be stable for contact annealing temperatures up to 550°C for 30 min.

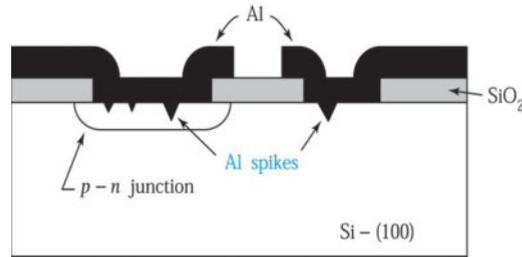


Fig. 24 Schematic view of aluminum films contacting silicon. Note the aluminum spiking in the silicon.

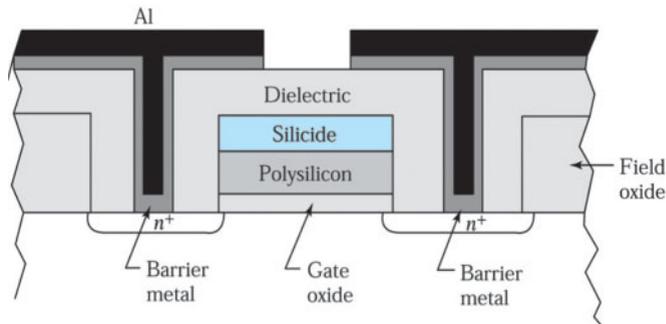


Fig. 25 Cross-sectional view of a MOSFET with a barrier metal between the aluminum and silicon and a composite gate electrode of silicide and polysilicon.

Electromigration

In Chapter 6 we discussed scaled-down devices. As the device becomes smaller, the corresponding current density becomes larger. High current densities can cause device failure due to electromigration, which refers to the transport of mass (i.e., atoms) in metals under the influence of current. It occurs by the transfer of momentum from the electrons to the positive metal ions. When a high current passes through thin metal conductors in integrated circuits, metal ions in some regions will pile up and voids will form in other regions. The pileup can short-circuit adjacent conductors, whereas the voids can result in an open circuit.

The mean time to failure (MTF) of a conductor due to electromigration can be related to the current density J and the activation energy E_a by

$$\text{MTF} \sim \frac{1}{J^2} \exp\left(\frac{E_a}{kT}\right). \quad (32)$$

Experimentally, a value of $E_a \cong 0.5$ eV is obtained for deposited aluminum. This indicates that low-temperature grain-boundary diffusion is the primary vehicle of material transport, since $E_a \cong 1.4$ eV would characterize the self-diffusion of single-crystal aluminum. The electromigration resistance of aluminum conductors can be increased by using several techniques, including alloying with copper (e.g., Al with 0.5% Cu), encapsulating the conductor in a dielectric, or incorporating oxygen during film deposition.

12.5.4 Copper Metallization

It is well known that both high-conductivity wiring and low-dielectric-constant insulators are required to lower the RC time delay of the interconnect network. Copper is the obvious choice for a new interconnection metallization because it has higher conductivity and higher electromigration resistance than aluminum. Copper can be deposited by PVD, CVD, and electrochemical methods. However, the use of Cu as an alternative material to Al in ULSI circuits has drawbacks, such as its tendency to corrode under standard chip manufacture conditions, its lack of a feasible dry-etching method or a stable self-passivating oxide similar to Al_2O_3 on Al, and its poor adhesion to dielectric materials, such as SiO_2 and low- k polymers. In this section, we discuss the copper metallization techniques.

Several different techniques for fabrication of multilevel Cu interconnects have been reported.^{18,19} The first method is a conventional method to pattern the metal lines followed by dielectric deposition. The second method is to pattern the dielectric layer first and fill copper metal into trenches. This step is followed by chemical mechanical polishing, discussed in Section 12.5.5, to remove the excess metal on the top surface of dielectric and leave Cu material in the vias and trenches. This method is also known as a damascene process.

Damascene Technology

The approach for fabricating a copper/low- k dielectric interconnect structure is by the “damascene” or “dual damascene” process. Figure 26 shows the dual damascene sequence for an advanced Cu interconnection structure. For a typical damascene structure, trenches for metal lines are defined and etched in the interlayer dielectric (ILD) and then followed by metal deposition of Ta(N)/Cu. The Ta(N) layer serves as a diffusion barrier layer and prevents copper from penetrating the low- k dielectric. The excess copper metal on the surface is removed to obtain a planar structure with metal inlays in the dielectric.

For the dual damascene process, the vias and trenches in the dielectric are defined using two lithography and reactive ion etching (RIE) steps before depositing the Ta(N)/Cu metal (Fig. 26a–c). Then a chemical-mechanical polishing process is used to remove the metal on the top surface, leaving the planarized wiring and via embedded in the insulator.²⁰ One special benefit of dual damascene is that the via plug is now of the same material as the metal line and the risk of via electromigration failure is reduced.

▶ EXAMPLE 6

If we replace Al with Cu wire associated with some low- k dielectric ($k = 2.6$) instead of SiO_2 layer, what percentage of reduction of RC time constant will be achieved? (The resistivity of Al is $2.7 \mu\Omega\text{-cm}$, and the resistivity of Cu is $1.7 \mu\Omega\text{-cm}$.)

SOLUTION

$$\frac{1.7}{2.7} \times \frac{2.6}{3.9} \times 100\% = 42\%.$$

12.5.5 Chemical-Mechanical Polishing

In recent years, the development of chemical-mechanical polishing (CMP) has become increasingly important for multilevel interconnection because it is the only technology that allows global planarization (i.e., making a flat surface across the whole wafer). It offers many advantages over other types of technologies—better global planarization over large or small structures, reduced defect density, and reduced plasma damage. Three CMP approaches are summarized in Table 4.

The CMP process consists of moving the sample surface against a pad that carries slurry between the sample surface and the pad. Abrasive particles in the slurry cause mechanical damage on the sample surface, loosening the material for enhanced chemical attack or fracturing off pieces of surface into the slurry where they dissolve or are swept away. The process is tailored to provide an enhanced material removal rate from high points on surfaces, thus affecting the planarization because most chemical actions are isotropic.

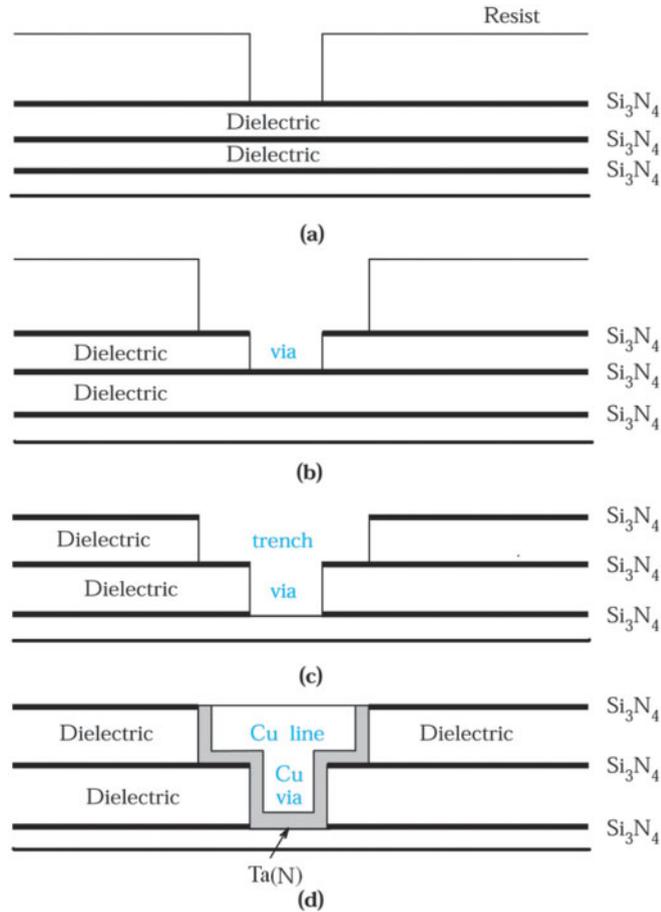


Fig. 26 Process sequence used to fabricate a Cu line-stud structure using dual damascene. (a) Resist stencil applied; (b) reactive ion etching dielectric and resist patterning; (c) trench and via definition; and (d) Cu deposition followed by chemical-mechanical polishing (CMP).

Mechanical grinding alone can theoretically achieve the desired planarization but is not desirable because of extensive associated damage to the material surfaces. There are three main parts of the process: the surface to be polished, the pad—the key medium enabling the transfer of mechanical action to the surface being polished—and the slurry, which provides both chemical and mechanical effects. Figure 27 shows the CMP setup.²¹

▶ EXAMPLE 7

The oxide removal rate and the removal rate of a layer underneath the oxide (called a stop layer) are $1r$ and $0.1r$, respectively. To remove $1\ \mu\text{m}$ oxide and $0.01\ \mu\text{m}$ stop layer, the total removal time is 5.5 min. Find the oxide removal rate.

SOLUTION

$$\frac{1}{1r} + \frac{0.01}{0.1r} = 5.5$$

$$\frac{1.1}{1r} = 5.5, \quad r = 0.2\ \mu\text{m}/\text{min.}$$

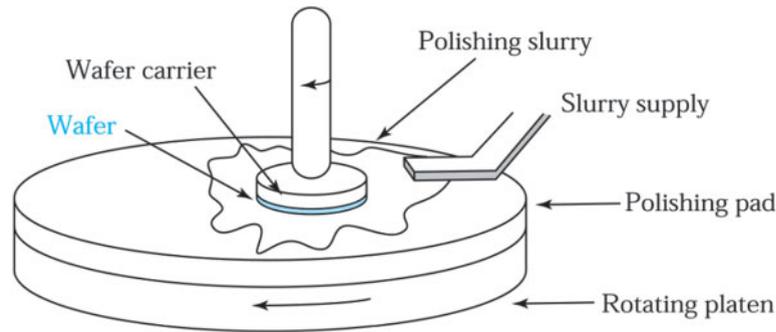


Fig. 27 A schematic of a CMP polisher

TABLE 4 THREE METHODS OF CHEMICAL-MECHANICAL POLISHING (CMP)

| Method | Wafer facing | Platen movement | Slurry feeding |
|-------------|--------------|--|-------------------------|
| Rotary CMP | Down | Rotary against rotating wafer carrier | Dripping to pad surface |
| Orbital CMP | Down | Orbital against rotating wafer carrier | Through the pad surface |
| Linear CMP | Down | Linear against rotating wafer carrier | Dripping to pad surface |

12.5.6 Silicide

Silicon forms many stable metallic and semiconducting compounds with metals called silicides. As the line width of these interconnections goes below 1 μm , the resistance of doped poly-Si (with a resistivity of the order of 500 $\mu\Omega\text{-cm}$) became unacceptable. Several metal silicides show low resistivity and high thermal stability, making them suitable for ULSI application. Silicides such as titanium silicide (TiSi_2), cobalt silicide (CoSi_2), and nickel silicide (NiSi) have reasonably low resistivities and are generally compatible with integrated-circuit processing. Silicides have become important metallization materials as devices have become smaller. One important application of silicide is for the MOSFET gate electrode, either alone or with doped polysilicon (polycide) above the gate oxide. In the following polycide and silicide processes, the presence of the poly-Si was necessary to keep the properties of the interface $\text{SiO}_2/\text{poly-Si}$ intact. Table 5 compares TiSi_2 , CoSi_2 , and NiSi .

Metal silicides have been used to reduce the contact resistance of the source and drain, the gate electrodes, and the interconnections. The self-aligned metal silicide technology (salicide) has been proven to be a highly attractive technique for improving the performance of submicron devices and circuits. The self-aligned process uses the silicide gate electrode as the mask to form the source and drain electrodes of a MOSFET (e.g., by ion implantation, considered in Chapter 14). This process can minimize the overlaps of these electrodes and thus reduce the parasitic capacitances.

TABLE 5 A COMPARISON OF TiSi_2 , CoSi_2 AND NiSi FILMS

| Properties | TiSi_2 | CoSi_2 | NiSi |
|---|----------------------|----------------------|-------------------|
| Resistivity ($\mu\Omega\text{-cm}$) | 13–16 | 15–20 | 10–20 |
| Silicide/metal ratio | 2.37 | 3.56 | 2.2 |
| Silicide/Si ratio | 1.04 | 0.97 | 1.2 |
| Reactive to native oxide | Yes | No | Yes |
| Silicidation temperature ($^\circ\text{C}$) | 800–850 | 550–900 | 400–550 |
| Film stress (dyne/cm^2) | 1.5×10^{10} | 1.2×10^{10} | 9.5×10^9 |

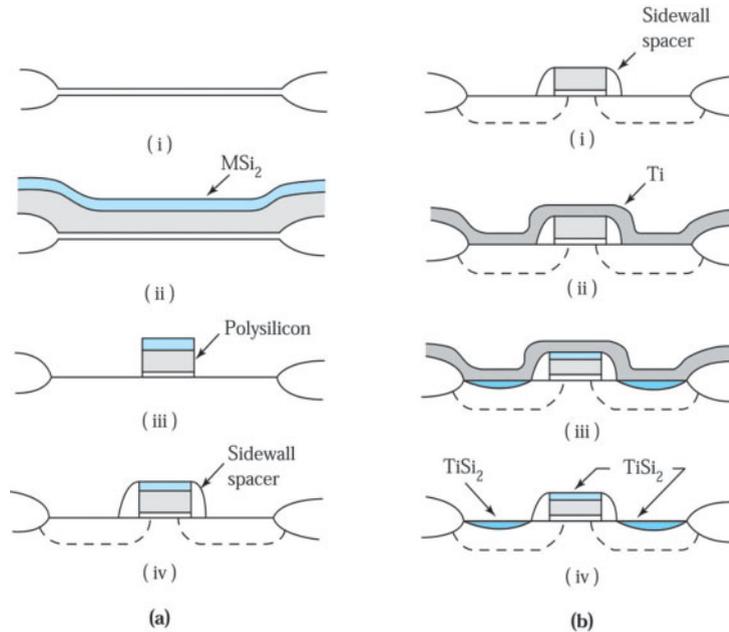


Fig. 28 Polycide and salicide processes. (a) Polycide structure: (i) gate oxide; (ii) polysilicon and silicide deposition, (iii) pattern polycide; and (iv) lightly doped drain (LDD) implant, sidewall formation, and *S/D* implant. (b) Salicide structure: (i) gate patterning (polysilicon only), LDD, sidewall, and *S/D* implant; (ii) metal (Ti, Co) deposition; (iii) anneal to form silicide; and (iv) selective (wet) etch to remove unreacted metal.

Figure 28 shows the polycide and salicide processes. A typical polycide formation sequence is shown in Fig. 28a. For sputter deposition, a high-temperature, high-purity compound target is used to ensure the quality of the silicide. The most commonly used silicides for the polycide process are WSi_2 , $TaSi_2$, and $MoSi_2$. They are all refractory, thermally stable, and resistant to processing chemicals. A self-aligned silicide process is illustrated in Fig. 28b. In the process, the polysilicon gate is patterned without any silicide, and a sidewall spacer (silicon oxide or silicon nitride) is formed to prevent shorting the gate to the source and drain during the silicidation process. A metal layer, either Ti or Co, is blanket-sputtered on the entire structure, followed by silicide sintering. Silicide is formed, in principle, only where the metal is in contact with Si. A wet chemical wash then rinses off the unreacted metal, leaving only the silicide. This technique eliminates the need to pattern the composite polycide gate structure and adds silicide to the source/drain area to reduce the contact resistance.

The silicides are promising materials for ULSI circuits because of their low resistivity and excellent thermal stability. Cobalt silicide has been widely investigated recently because of its lowest resistivity and high-temperature thermal stability. However, cobalt is sensitive to native oxide as well as an oxygen-contained environment, and a large amount of silicon is consumed during silicidation.

From the polycide and salicide formation processes in Figure 28, we can deduce the following desired properties for a silicide material:

1. Low resistivity (limits contact alignment issues for multilayers and reduces device resistance).
2. Etch selectivity of the silicide vs. the metal (allows self-aligned process).
3. Etch resistance in reactive ion etch (RIE) environment (allows opening of via holes).
4. Acceptable diffusion barrier properties.
5. Low roughness (gives a minimal junction penetration).
6. Preferably high resistance to oxidation.

Beside these six characteristics, the silicide must also fill the following criteria: high morphological stability, minimal Si consumption (limited doped Si available), and controlled film stress.

► SUMMARY

Modern semiconductor device fabrication requires the use of thin films. Currently, there are four important types of films—thermal oxides, dielectric layers, poly-crystalline silicon, and metal films. The major issues related to film formation are low-temperature processing, step coverage, selective deposition, uniformity, film quality, planarization, throughput, and large-wafer capacity.

Thermal oxidation offers the best quality for the Si-SiO₂ interface and has the lowest interface trap density. Therefore, it is used to form the gate oxide and the field oxide. LPCVD of dielectrics and polysilicon offers conformal step coverage. In contrast, PVD and atmospheric-pressure CVD generally result in nonconformal step coverage. CMP offers global planarization and reduces defect density. Conformal step coverage and planarization are also required for precise pattern transfer at lithography level below 100 nm. Pattern transfer technology is discussed in the next chapter. An emerging technology for film formation is atomic-layer deposition, which is capable of depositing oxide and metal thin films of the order of a monolayer in thickness.

To minimize the *RC* time delay due to parasitic resistance and capacitance, the silicide process for ohmic contacts, copper metallization for interconnects, and low-dielectric-constant materials for interlayer films are used extensively to meet the requirements for the multilevel interconnect structures of ULSI circuits. In addition, we have investigated high-dielectric-constant materials to improve gate insulator performance and to increase the capacitance per unit area for DRAM.

► REFERENCES

1. E. H. Nicollian and J. R. Brews, *MOS Physics and Technology*, Wiley, New York, 1982.
2. B. E. Deal and A. S. Grove, "General Relationship for the Thermal Oxidation of Silicon," *J. Appl. Phys.*, **36**, 3770 (1965).
3. J. D. Meindl et al., "Silicon Epitaxy and Oxidation," in F. Van de Wiele, W. L. Engl, and P. O. Jespers, Eds., *Process and Device Modeling for Integrated Circuit Design*, Noorhoff, Leyden, 1977.
4. For a discussion of film deposition, see, for example, A.C. Adams, "Dielectric and Polysilicon Film Deposition," in S. M. Sze, Ed., *VLSI Technology*, McGraw-Hill, New York, 1983.
5. K. Eujino et al., "Doped Silicon Oxide Deposition by Atmospheric Pressure and Low Temperature Chemical Vapor Deposition Using Tetraethoxysilane and Ozone," *J. Electrochem. Soc.*, **138**, 3019 (1991).
6. A. C. Adams and C. D. Capio, "Planarization of Phosphorus-Doped Silicon Dioxide," *J. Electrochem. Soc.*, **127**, 2222 (1980).
7. T. Yamamoto et al., "An Advanced 2.5nm Oxidized Nitride Gate Dielectric for Highly Reliable 0.25 μm MOSFETs," *Symp. on VLSI Technol. Dig. of Tech. Pap.*, 1997, p. 45.
8. K. Kumar, et al., "Optimization of Some 3 nm Gate Dielectrics Grown by Rapid Thermal Oxidation in a Nitric Oxide Ambient," *Appl. Phys. Lett.*, **70**, 384 (1997).
9. T. Homma, "Low Dielectric Constant Materials and Methods for Interlayer Dielectric Films in Ultralarge-Scale Integrated Circuit Multilevel Interconnects," *Mater. Sci. Eng.*, **23**, 243 (1998).
10. H. N. Yu et al., "1 μm MOSFET VLSI Technology. Part I—An Overview," *IEEE Trans. Electron Devices*, **ED-26**, 318 (1979).
11. J. M. Andrews, "Electrical Conduction in Implanted Polycrystalline Silicon," *J. Electron. Mater.*, **8**, 3, 227 (1979).
12. R. Doering and Y. Nishi, Eds., *Handbook of Semiconductor Manufacturing Technology*, 2nd Ed., CRC Press, FL, 2008.
13. M. J. Buiting, A. F. Otterloo, and A. H. Montree, "Kinetic aspects of the LPCVD of titanium nitride from titanium tetrachloride and ammonia," *J. Electrochem. Soc.*, **138**, 500 (1991).

14. R. Tobe, et al., "Plasma-Enhanced CVD of TiN and Ti Using Low-Pressure and High-Density Helicon Plasma," *Thin Solid Film*, **281–282**, 155 (1996).
15. J. Hu, et al., "Electrical Properties of Ti/TiN Films Prepared by Chemical Vapor Deposition and Their Applications in Submicron Structures as Contact and Barrier Materials," *Thin Solid Film*, **308**, 589 (1997).
16. M. Hansen and A. Anderko, *Constitution of Binary Alloys*, McGraw-Hill, New York, 1958.
17. D. Pramanik and A. N. Saxena, "VLSI Metallization Using Aluminum and Its Alloys," *Solid State Tech.*, **26**, No. 1, 127 (1983), **26**, No. 3, 131 (1983).
18. C. L. Hu and J. M. E. Harper, "Copper Interconnections and Reliability," *Matter. Chem. Phys.*, **52**, 5 (1998).
19. P. C. Andricacos et al., "Damascene Copper Electroplating for Chip Interconnects," *193rd Meet. Electrochem. Soc.*, 1998, p. 3.
20. J. M. Steigerwald et al., "Chemical Mechanical Planarization of Microelectronic Materials," Wiley, New York, 1997.
21. L. M. Cook et al., *Theoretical and Practical Aspects of Dielectric and Metal CMP*, Semicond. Int., p. 141 (1995).

► **PROBLEMS (* DENOTES DIFFICULT PROBLEMS)**

FOR SECTION 12.1 THERMAL OXIDATION

1. A p -type $\langle 100 \rangle$ -oriented silicon wafer with a resistivity of $10 \Omega\text{-cm}$ is placed in a wet oxidation system to grow a field oxide of $0.45 \mu\text{m}$ at 1050°C . Determine the time required to grow the oxide.
- *2. After the first oxidation as given in Prob. 1, a window is opened in the oxide to grow a gate oxide at 1000°C for 20 minutes in dry oxidation. Find the thicknesses of the gate oxide and the total field oxide.
3. Show that Eq. 11 reduces to $x^2 = Bt$ for long times and to $x = B/A(t + \tau)$ for short times.
4. Determine the diffusion coefficient D for dry oxidation of $\langle 100 \rangle$ -oriented silicon samples at 980°C and 1 atm.

FOR SECTION 12.2 CHEMICAL VAPOR DEPOSITION OF DIELECTRICS

5. In a plasma-deposited silicon nitride that contains 20 at% hydrogen and has a silicon-to-nitrogen ratio (Si/N) of 1.2, find x and y in the empirical formula of SiN_xH_y . (b) If the variation of film resistivity with Si/N ratio is given by $5 \times 10^{28} \exp(-33.3\gamma)$ for $2 > \gamma > 0.8$, where γ is the ratio, find the resistivity of the film in (a).
6. The dielectric constants of SiO_2 , Si_3N_4 , and Ta_2O_5 are about 3.9, 7.6, and 25, respectively. What is the capacitance ratio for the capacitors with the Ta_2O_5 and oxide/nitride/oxide dielectrics for the same dielectric thickness, provided the oxide/nitride/oxide has thickness ratio 1:1:1 for the oxide to the nitride?
7. In Prob. 6, if BST with a dielectric constant of 500 is chosen to replace Ta_2O_5 , calculate the area reduction ratio to maintain the same capacitance if the two films have the same thickness.
8. In Prob. 6, calculate the equivalent thickness of the Ta_2O_5 in terms of SiO_2 thickness if both have the same capacitance. Assume the actual thickness of Ta_2O_5 is $3t$.
9. In a silane-oxygen reaction to deposit undoped SiO_2 film, the deposition rate is 15 nm/min at 425°C . What temperature is required to double the deposition rate?

10. The P-glass flow process requires temperatures above 1000°C. As device dimensions become smaller in ULSI, we must use lower temperatures. Suggest methods to obtain a smooth topography at < 900°C for deposited silicon dioxide that can be used as an insulator between metal layers.

FOR SECTION 12.3 CHEMICAL VAPOR DEPOSITION OF POLYSILICON

11. Why is silane more often used for polysilicon deposition than silicon chloride?
 12. Explain why the deposition temperature for polysilicon films is moderately low, usually between 600°–650°C.

FOR SECTION 12.4 ATOMIC LAYER DEPOSITION

13. In an ideal case, calculate the surface density of Al₂O₃ by ALD? (The density of Al₂O₃ is 3 g/cm³.)
 14. Calculate the deposition rate of Al₂O₃ by ALD using Al(CH₃)₃ and water as precursors, if ALD is operated under saturation conditions. The expose time of each precursor is 1 second and the purge time for each precursor is 1 second.

FOR SECTION 12.5 METALLIZATION

15. An e-beam evaporation system is used to deposit aluminum to form MOS capacitors. If the flatband voltage of the capacitance is shifted by 0.5 V because of e-beam radiation, find the number of fixed oxide charges (the silicon dioxide thickness is 50 nm). How can these charges be removed?
 16. A metal line ($L = 20 \mu\text{m}$, $W = 0.25 \mu\text{m}$) has a sheet resistance 5 Ω/sq. Calculate the resistance of the metal line.
 17. Calculate the thickness of the TiSi₂ and CoSi₂, where the initial Ti and Co film thicknesses are 30 nm.
 18. Compare the advantages and disadvantages of TiSi₂ and CoSi₂ for salicide applications.
 19. A dielectric material is placed between the two parallel metal lines. The length $L = 1 \text{ cm}$, width $W = 0.28 \mu\text{m}$, thickness $T = 0.3 \mu\text{m}$, and spacing $S = 0.36 \mu\text{m}$. (a) Calculate the RC time delay. The metal is Al with a resistivity of 2.67 μΩ-cm and the dielectric is oxide with dielectric constant 3.9. (b) Calculate the RC time delay. The metal is Cu with a resistivity of 1.7 μΩ-cm and the dielectric is organic polymer with dielectric constant 2.8. (c) Compare the results in (a) and (b). How much can we decrease the RC time delay?
 20. Repeat Prob. 19 (a) and (b) if the fringing factor for the capacitors is 3. The fringing factor is due to the spreading of the electric-field lines beyond the length and width of the metal lines.
 *21. To avoid electromigration problems, the maximum allowed current density in an aluminum runner is about $5 \times 10^5 \text{ A/cm}^2$. If the runner is 2 mm long, 1 μm wide, and nominally 1 μm thick, and if 20% of the runner length passes over steps and is only 0.5 μm thick there, find the total resistance of the runner if the resistivity is $3 \times 10^{-6} \Omega\text{-cm}$. Find the maximum voltage that can be applied across the runner.
 22. To use Cu for wiring one must overcome several obstacles: the diffusion of Cu through SiO₂, adhesion of Cu to SiO₂, and corrosion of Cu. One way to overcome these obstacles is to use a cladding/adhesion layer (e.g., Ta or TiN) to protect the Cu wires. Consider a clad Cu wire with a square cross section of $0.5 \mu\text{m} \times 0.5 \mu\text{m}$, and compare it with a layered TiN/Al/TiN wire of the same size, with the top and bottom TiN layers 40 nm and 60 nm thick, respectively. What is the maximum thickness of the cladding layer if the resistance of the clad Cu wire and the TiN/Al/TiN wire is the same?

Lithography and Etching

- ▶ 13.1 OPTICAL LITHOGRAPHY
 - ▶ 13.2 NEXT-GENERATION LITHOGRAPHIC METHODS
 - ▶ 13.3 WET CHEMICAL ETCHING
 - ▶ 13.4 DRY ETCHING
 - ▶ SUMMARY
-

Lithography is the process of transferring patterns of geometric shapes on a mask to a thin layer of radiation-sensitive material (called resist) covering the surface of a semiconductor wafer.¹ These patterns define the various regions in an integrated circuit such as the implantation regions, the contact windows, and the bonding-pad areas. The resist patterns defined by the lithographic process are not permanent elements of the final device but only replicas of circuit features. To produce circuit features, these resist patterns must be transferred once more into the underlying layers comprising the device. The pattern transfer is accomplished by an etching process that selectively removes unmasked portions of a layer.²

Specifically, we cover the following topics:

- The importance of a clean room for lithography.
- The most widely used lithographic method—optical lithography and its resolution-enhancement techniques.
- Advantages and limitations of other lithographic methods.
- Mechanisms for wet chemical etching of semiconductors, insulators, and metal films.
- Dry etching (also called plasma-assisted etching) for high-fidelity pattern transfer.

▶ 13.1 OPTICAL LITHOGRAPHY

The vast majority of lithographic equipment for integrated-circuit (IC) fabrication is optical equipment using ultraviolet light ($\lambda \cong 0.2\text{--}0.4\ \mu\text{m}$). In this section we consider the exposure tools, the masks, the resists, and resolution-enhancement techniques used for optical lithography. We also consider the pattern transfer process, which serves as a basis for other lithographic systems. We first briefly consider the *clean room*, because all lithographic processes must be performed in an ultraclean environment.

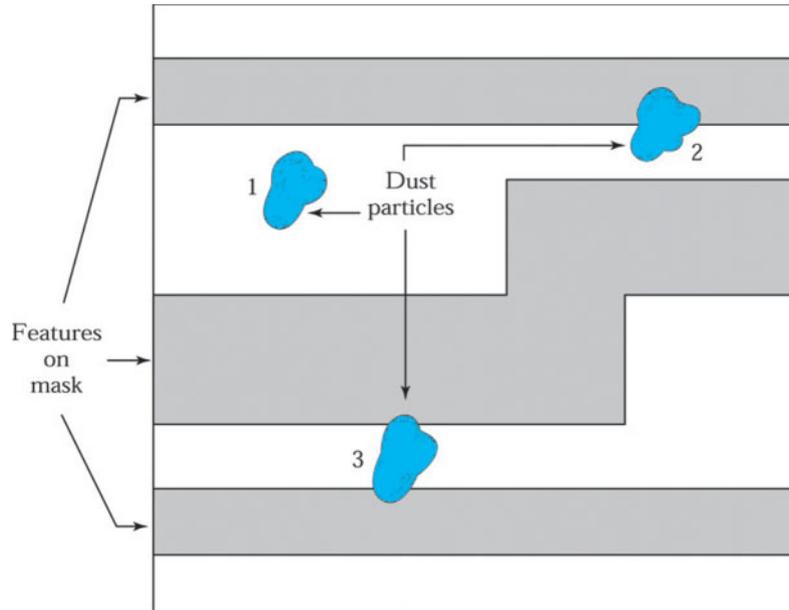


Fig. 1 Various ways in which dust particles can interfere with photomask patterns.³

13.1.1 The Clean Room

An IC fabrication facility requires a clean processing room, especially in the area used for lithography. The need for such a clean room arises because dust particles in the air can settle on semiconductor wafers and lithographic masks and can cause defects in the devices and hence circuit failure. For example, a dust particle on a semiconductor surface can disrupt the single-crystal growth of an epitaxial film, causing the formation of dislocations. A dust particle incorporated into the gate oxide can result in enhanced conductivity and cause device failure due to low breakdown voltage. The situation is even more critical in the lithographic area. When dust particles adhere to the surface of a photomask, they behave as opaque patterns on the mask, and these patterns will be transferred to the underlying layer along with the circuit patterns on the mask. Figure 1 shows three dust particles on a photomask.³ Particle 1 may result in the formation of a pinhole in the underlying layer. Particle 2 is located near a pattern edge and may cause a constriction of current flow in a metal runner. Particle 3 can lead to a short circuit between the two conducting regions and render the circuit useless.

In a clean room, the total number of dust particles per unit volume must be tightly controlled along with the temperature and humidity. Figure 2 shows the particle-size distribution curves for various *classes* of clean rooms. We have two systems to define the classes of clean room.⁴ In the English system, the numerical designation of the class is taken from the maximum allowable number of particles $0.5\ \mu\text{m}$ and larger, per *cubic foot*. For the metric system, the class is taken from the logarithm (base 10) of the maximum allowable number of particles $0.5\ \mu\text{m}$ and larger, per *cubic meter*. For example, a class 100 clean room (English system) has a dust count of 100 particles/ ft^3 with particle diameters of $0.5\ \mu\text{m}$ and larger, whereas a class M 3.5 clean room (metric system) has a dust count of $10^{3.5}$ or about 3500 particles/ m^3 with particle diameters of $0.5\ \mu\text{m}$ and larger. Since $100\ \text{particles}/\text{ft}^3 = 3500\ \text{particles}/\text{m}^3$, a class 100 in English system corresponds to a class M 3.5 in the metric system.

For most IC fabrication areas, a class 100 clean room is required, that is, the dust count must be about four orders of magnitude below that of ordinary room air. However, for the lithography area, a class 10 clean room or one with a lower dust count is required.

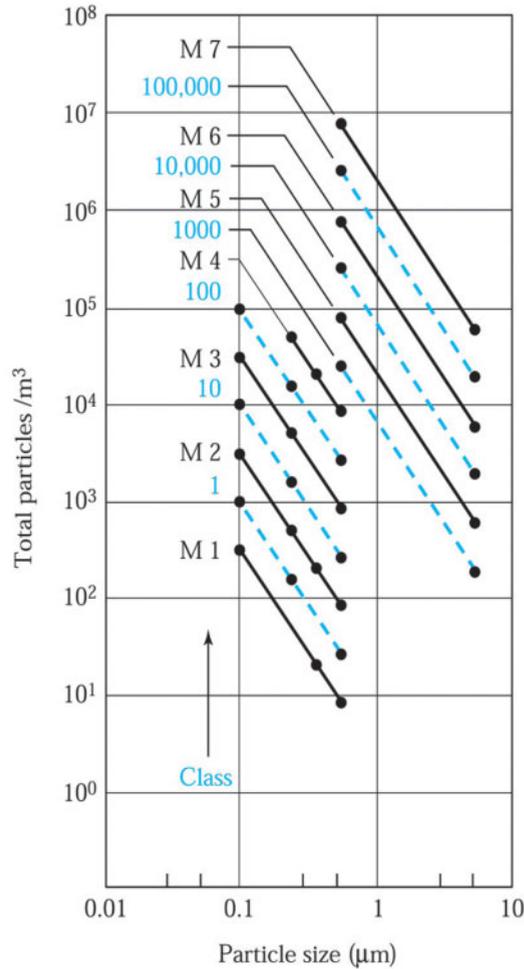


Fig. 2 Particle-size distribution curve for English (---) and metric (—) classes of clean rooms.⁴

► EXAMPLE 1

If we expose a 300-mm wafer for 1 minute to an air stream under a laminar-flow condition at 30 m/min, how many dust particles will land on the wafer in a class-10 clean room?

SOLUTION For a class 10 clean room, there are 350 particles (0.5 µm and larger) per cubic meter. The air volume that goes over the wafer in 1 minute is

$$30 \text{ m/min} \times \pi \left(\frac{0.3 \text{ m}}{2} \right)^2 \times 1 \text{ minute} = 2.12 \text{ m}^3.$$

The number of dust particles (0.5 µm and larger) contained in the air volume is $350 \times 2.12 = 742$ particles. Therefore, if there are 800 IC chips on the wafer, the particle count amounts to one particle on each of 92% of the chips. Fortunately, only a fraction of the particles that land adhere to the wafer surface, and of those only a fraction are at a circuit location critical enough to cause a failure. However, the calculation indicates the importance of the clean room. ◀

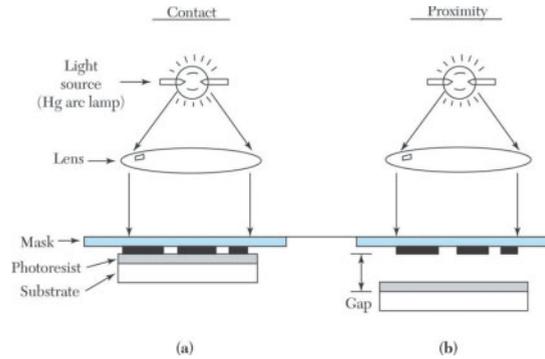


Fig. 3 Schematic of optical shadow printing techniques:¹ (a) contact printing, (b) proximity printing.

13.1.2 Exposure Equipment

The pattern transfer process is accomplished by using lithographic exposure equipment. The performance of exposure equipment is determined by three parameters: resolution, registration, and throughput. *Resolution* is the minimum feature dimension that can be transferred with high fidelity to a resist film on a semiconductor wafer. *Registration* is a measure of how accurately patterns on successive masks can be aligned (or overlaid) on previously defined patterns on the wafer. *Throughput* is the number of wafers that can be exposed per hour for a given mask level.

There are basically two optical exposure methods: shadow printing and projection printing.^{5,6} Shadow printing may have the mask and wafer in direct contact with each other as in *contact printing*, or in close proximity as in *proximity printing*. Figure 3a shows a basic setup for contact printing where a resist-coated wafer is brought into physical contact with a mask, and the resist is exposed by a nearly collimated beam of ultraviolet light through the back of the mask for a fixed time. The intimate contact between resists and mask provides a resolution of $\sim 1 \mu\text{m}$. However, contact printing suffers a major drawback caused by dust particles. A dust particle or a speck of silicon dust on the wafer can be imbedded into the mask when the mask makes contact with the wafer. The imbedded particle causes permanent damage to the mask and results in defects in the wafer with each succeeding exposure.

To minimize mask damage, the proximity exposure method is used. Figure 3b shows the basic setup. It is similar to the contact printing method except that there is a small gap (10–50 μm) between the wafer and the mask during exposure. The small gap results in optical diffraction at feature edges on the photomask: that is, when light passes by the edges of an opaque mask feature, fringes are formed and some light penetrates into the shadow region. As a result, the resolution is degraded to the 2–5 μm range.

In shadow printing, the minimum linewidth [or critical dimension (CD)] that can be printed is roughly

$$\text{CD} \cong \sqrt{\lambda g}, \quad (1)$$

where λ is the wavelength of the exposure radiation and g is the gap between the mask and wafer and includes the thickness of the resist. For $\lambda = 0.4 \mu\text{m}$ and $g = 50 \mu\text{m}$, the CD is 4.5 μm . If we reduce λ to 0.25 μm (the wavelength range of 0.2–0.3 μm is in the deep-UV spectral region) and g to 15 μm , CD becomes 2 μm . Thus, there is an advantage in reducing both λ and g . However, for a given distance g , any dust particle with a diameter larger than g can potentially cause mask damage.

To avoid the mask damage problem associated with shadow printing, projection-printing exposure equipment has been developed to project an image of the mask patterns onto a resist-coated wafer many centimeters away from the mask. To increase resolution, only a small portion of the mask is exposed at a time, allowing a uniform source of light. The small image area is scanned or stepped over the wafer to cover the entire wafer surface. Figure 4a shows a 1:1 wafer scan projection system.^{6,7} A narrow, arc-shaped image field $\sim 1 \text{ mm}$ in width serially transfers the slit image of the mask onto the wafer. The image size on the wafer is the same as on the mask.

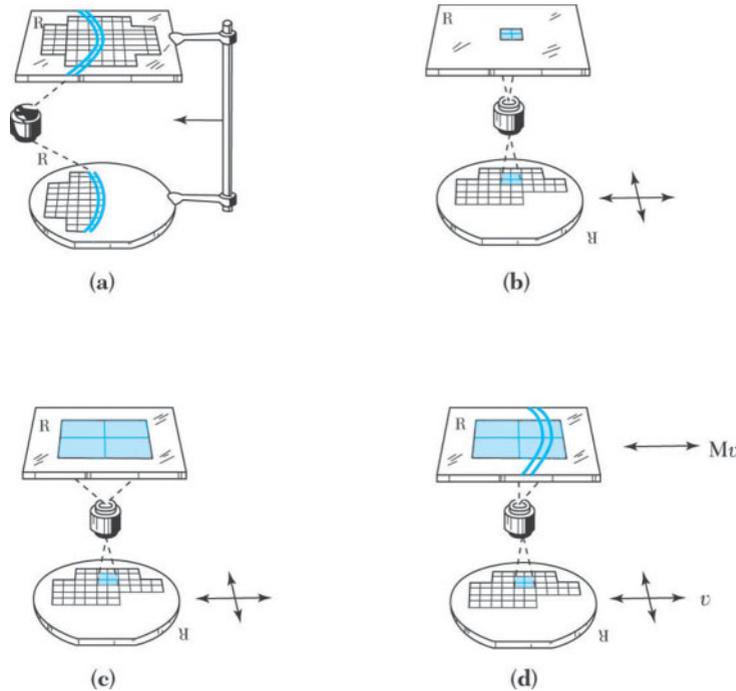


Fig. 4 Image-partitioning techniques for projection printing: (a) annual-field wafer scan, (b) 1:1 step-and-repeat, (c) M :1 reduction step-and-repeat, and (d) M :1 reduction step-and-scan.^{6,7}

The small image field can also be stepped over the surface of the wafer by two-dimensional translations of the wafer only, while the mask remains stationary. After the exposure of one chip site, the wafer is moved to the next chip site and the process is repeated. Figure 4b and 4c show the partitioning of the wafer image by *step-and-repeat projection* with a ratio of 1:1 or at a demagnification ratio M :1 (e.g., 10:1 for a 10 times reduction on the wafer), respectively. The demagnification ratio is an important factor in our ability to produce both the lens and the mask from which we wish to print. The 1:1 optical systems are easier to design and fabricate than 10:1 or 5:1 reduction systems, but it is much more difficult to produce defect-free masks at 1:1 than at a 10:1 or a 5:1 demagnification ratio.

Reduction projection lithography can also print larger wafers without redesigning the stepper lens as long as the field size (i.e., the exposure area onto the wafer per se) of the lens is large enough to contain one or more IC chips. When the chip size exceeds the field size of the lens, further partitioning of the image on the mask is necessary. In Fig. 4d the image field on the mask can be a narrow, arc-shaped for M :1 step-and-scan projection lithography. For the step-and-scan system, we have two-dimensional translations of the wafer with speed v , and one-dimensional translation of the mask with M times that of the wafer speed.

The resolution of a projection system l_m is usually determined by the quality of the lens, but is ultimately limited by diffraction and given by

$$l_m = k_1 \frac{\lambda}{\text{NA}}, \quad (2)$$

where λ is again the exposure wavelength, k_1 is a process-dependent factor, and NA is the numerical aperture, which is given by

$$\text{NA} = \bar{n} \sin \theta = \bar{n} \sin (\tan^{-1} D/2f) \approx \bar{n} (D/2f) \quad (3)$$

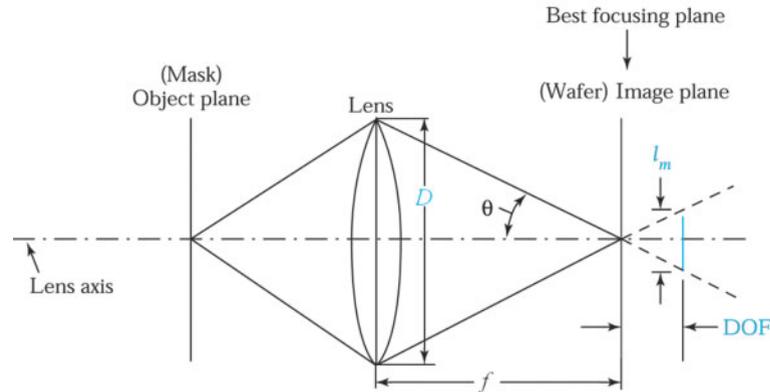


Fig. 5 Simple image system.⁵

with \bar{n} the index of refraction in the image medium (usually air, where $\bar{n} = 1$), θ the half-angle of the cone of light converging to a point image at the wafer, D the diameter of the lens and f the focal length, as shown⁵ in Fig. 5. Hence the numerical aperture of an optical system is a dimensionless number that characterizes the range of angles over which the system can accept or emit light. Also shown in the figure is the depth of focus (DOF), which can be expressed as

$$\text{DOF} = \frac{\pm l_m / 2}{\tan \theta} \approx \frac{\pm l_m / 2}{\sin \theta} = k_2 \frac{\lambda}{(\text{NA})^2}, \quad (4)$$

where k_2 is another process-dependent factor. The depth of focus is a measure of the distance from the lens in which the film or sensor plane will remain in focus. In photolithography, it is useful to specify the flatness and thickness of the resist to assure sharp focus.

Equation 2 indicates that resolution can be improved (i.e., smaller l_m) by either reducing the wavelength or increasing NA or both. However, Eq. 4 indicates that the DOF degrades much more rapidly by increasing NA than by decreasing λ . This explains the trend toward shorter-wavelength sources in optical lithography.

The high-pressure mercury-arc lamp is widely used in exposure equipment because of its high intensity and reliability. The mercury-arc spectrum is composed of several peaks. The terms G-line, H-line, and I-line refer to the peaks at 436 nm, 405 nm, and 365 nm, respectively. I-line lithography with 5:1 step-and-repeat projection can offer a resolution of 0.3 μm with resolution enhancement techniques (see Section 13.1.6). Advanced exposure equipment such as the 248 nm lithographic system using a KrF excimer laser, the 193 nm lithographic system using an ArF excimer laser, and the immersion 193 nm system (in which the lens is immersed in water to increase the index of refraction from 1 to 1.33) have been developed for mass production with resolutions of 180 nm, 100 nm, and below 70 nm, respectively.

13.1.3 Masks

Reduction techniques are usually used to fabricate masks for IC manufacturing. The first step in mask making is to use a computer-aided design (CAD) system in which designers can completely describe the circuit patterns electrically. The digital data produced by the CAD system then drives a pattern generator, which is an electron-beam lithographic system (see Section 13.2.1) that transfers the patterns directly to electron-sensitized mask. The mask consists of a fused silica substrate covered with a chromium layer. The circuit pattern is first transferred to the electron-sensitized layer (electron resist), which is transferred once more into the underlying chromium layer for the finished mask. The details of pattern transfer are considered in Section 13.1.5.

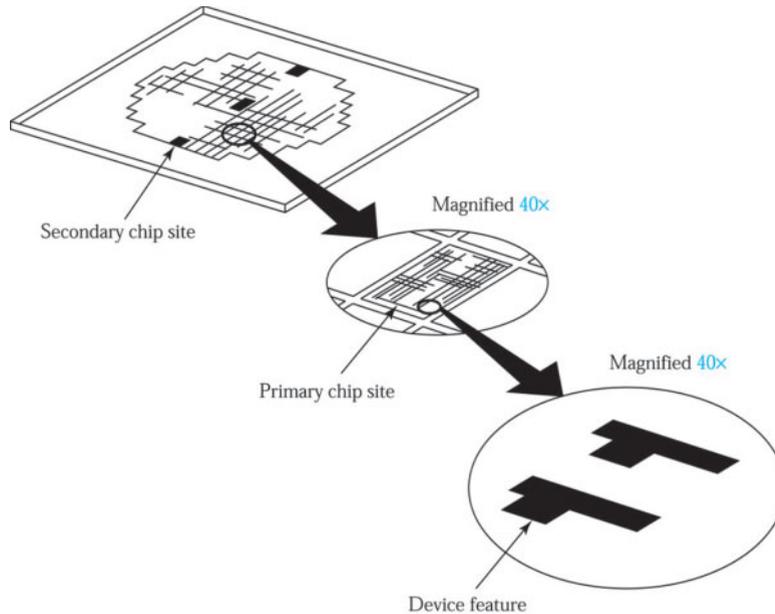


Fig. 6 An integrated-circuit photomask.¹

The patterns on a mask represent one level of an IC design. The composite layout is broken into mask levels that correspond to the IC process sequence such as the isolation region on one level, the gate region on another, and so on. Typically, 15–20 different mask levels are required for a complete IC process cycle.

The standard-size mask substrate is a fused silica plate 15×15 cm square and 0.6 cm thick. The size is needed to accommodate the lens field sizes for 4:1 or 5:1 optical exposure equipment, whereas the thickness is required to minimize pattern placement errors due to substrate distortion. The fused silica plate is needed for its low coefficient of thermal expansion, high transmission at shorter wavelengths, and mechanical strength. Figure 6 shows a mask on which patterns of geometric shapes have been formed. A few secondary-chip sites used for process evaluation are also included in the mask.

One of the major concerns about masks is the defect density. Mask defects can be introduced during the manufacture of the mask or during subsequent lithographic processes. Even a small mask-defect density has a profound effect on the final IC yield. The *yield* is defined as the ratio of good chips per wafer to the total number of chips per wafer. As a first-order approximation, the yield Y for a given masking level can be expressed as

$$Y \cong e^{-DA}, \quad (5)$$

where D is the average number of “fatal” defects per unit area and A is the area of an IC chip. If D remains the same for all mask levels (e.g., $N = 10$ levels), then the final yield becomes

$$Y \cong e^{-NDA}. \quad (6)$$

Figure 7 shows the mask-limit yield for a 10-level lithographic process as a function of chip size for various values of defect densities. For example, for $D = 0.25$ defect/cm², the yield is 10% for a chip size of 90 mm², and it drops to about 1% for a chip size of 180 mm². Therefore, inspection and cleaning of masks are important to achieve high yields on large chips. Of course, an ultraclean processing area is mandatory for lithographic processing.

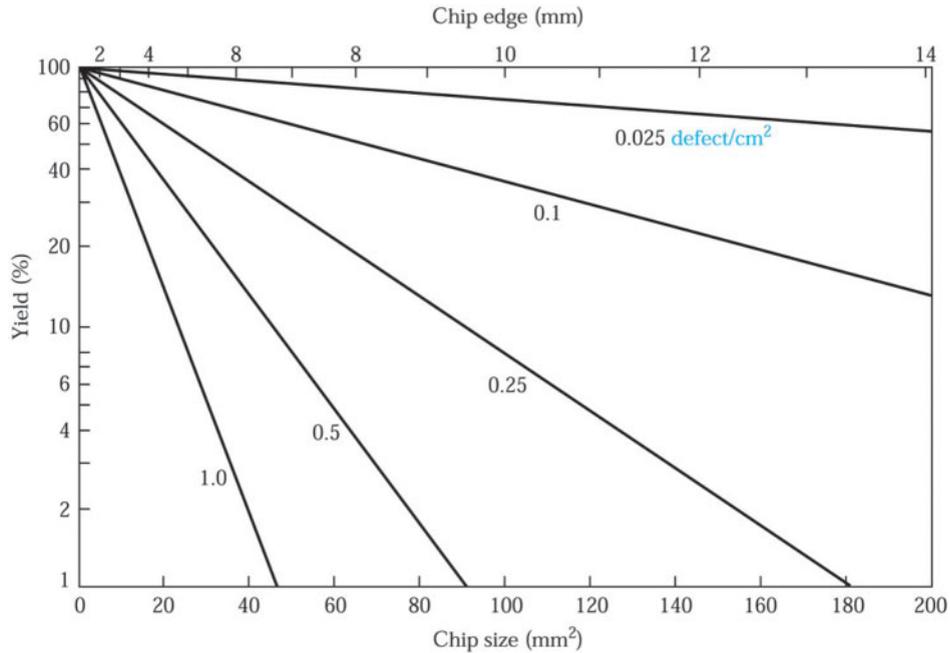


Fig. 7 Yield for a 10-mask lithographic process with various defect densities per level.

13.1.4 Photoresist

The photoresist is a radiation-sensitive compound. Photoresists are classified as positive and negative, depending on how they respond to radiation. For positive resists, the exposed regions become more soluble and thus more easily removed in the development process. The net result is that the patterns formed (also called images) in the positive resist are the same as those on the mask. For negative resists, the exposed regions become less soluble, and the patterns formed in the negative resist are the reverse of the mask patterns.

Positive photoresists have three components: a photosensitive compound, a base resin, and an organic solvent. Prior to exposure, the photosensitive compound is insoluble in the developer solution. After exposure, the photosensitive compound absorbs radiation in the exposed pattern areas, changes its chemical structure, and becomes soluble in the developer solution. After development, the exposed areas are removed.

Negative photoresists are polymers combined with a photosensitive compound. After exposure, the photosensitive compound absorbs the optical energy and converts it into chemical energy to initiate a polymer linking reaction. This reaction causes cross linking of the polymer molecules. The cross-linked polymer has a higher molecular weight and becomes insoluble in the developer solution. After development, the unexposed areas are removed. One major drawback of a negative photoresist is that in the development process, the whole resist mass swells by absorbing developer solvent. This swelling action limits the resolution of negative photoresists.

Figure 8a shows a typical exposure response curve and image cross section for a positive resist.¹ The response curve describes the percentage of resist remaining after exposure and development versus the exposure energy. Note that the resist has a finite solubility in its developer, even without exposure to radiation. As the exposure energy increases, the solubility gradually increases until at a threshold energy E_T , the resist becomes completely soluble. The sensitivity of a positive resist is defined as the energy required to produce complete solubility in the exposed region. Thus, E_T corresponds to the sensitivity. In addition to E_T , a parameter γ , the contrast ratio, is defined to characterize the resist:

$$\gamma \equiv \left[\ln \left(\frac{E_T}{E_l} \right) \right]^{-1}, \quad (7)$$

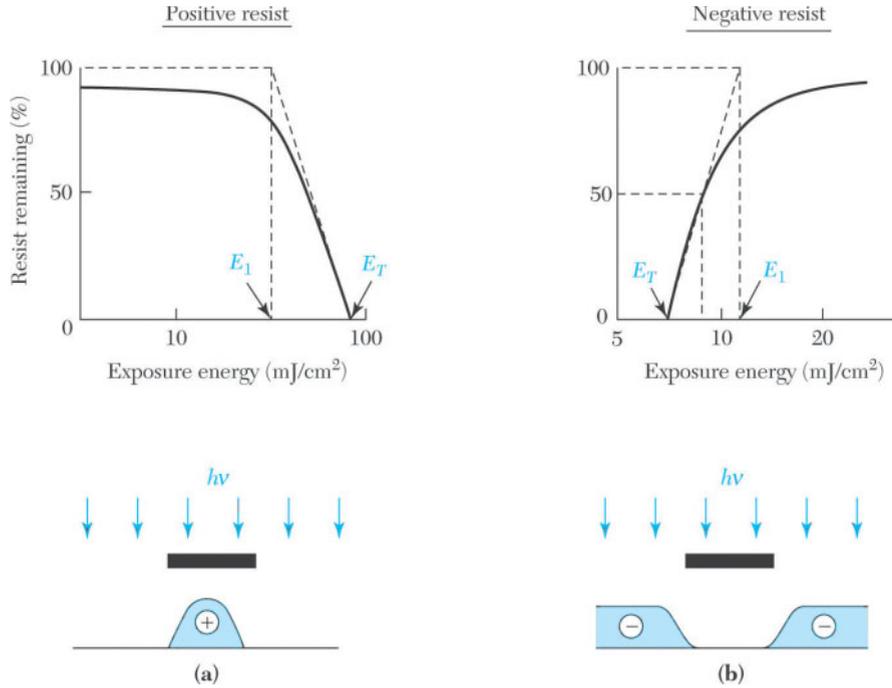


Fig. 8 Exposure-response curve and cross section of the resist image after development.¹ (a) Positive photoresist; (b) negative photoresist.

where E_1 is the energy obtained by drawing the tangent at E_T to reach 100% resist thickness as shown in Fig. 8a. A larger γ implies a higher solubility of the resist with an incremental increase of exposure energy and yields sharper images.

The image cross section in Fig. 8a illustrates the relationship between the edges of a photomask image and the corresponding edges of the resist images after development. The edges of the resist image are generally not at the vertically projected positions of the mask edges because of *diffraction*. The edge of the resist image corresponds to the position where the total absorbed optical energy equals the threshold energy E_T .

Figure 8b shows the exposure-response curve and image cross section for a negative resist. The negative resist remains completely soluble in the developer solution for exposure energies lower than the threshold energy E_T . Above E_T , more of the resist film remains after development. At exposure energies twice the threshold energy, the resist film becomes essentially insoluble in the developer. The sensitivity of a negative resist is defined as the energy required to retain 50% of the original resist film thickness in the exposed region. The parameter γ is defined similarly to γ in Eq. 7 except that E_1 and E_T are interchanged. The image cross section for the negative resist (Fig. 8b) is also influenced by the diffraction effect.

► EXAMPLE 2

Find the parameter γ for the photoresists shown in Fig. 8.

SOLUTION For the positive resist, $E_T = 90 \text{ mJ/cm}^2$ and $E_1 = 45 \text{ mJ/cm}^2$:

$$\gamma = \left[\ln \left(\frac{E_T}{E_1} \right) \right]^{-1} = \left[\ln \left(\frac{90}{45} \right) \right]^{-1} = 1.4.$$

For the negative resist, $E_T = 7 \text{ mJ/cm}^2$ and $E_1 = 12 \text{ mJ/cm}^2$:

$$\gamma = \left[\ln \left(\frac{E_1}{E_T} \right) \right]^{-1} = \left[\ln \left(\frac{12}{7} \right) \right]^{-1} = 1.9. \quad \blacktriangleleft$$

For deep UV lithography (e.g., 248 and 193 nm), we cannot use conventional photoresists because these resists require a high-dose exposure in deep UV, which will cause lens damage and lower throughput. The chemical-amplified resist (CAR) has been developed for the deep UV process. CAR consists of a photo-acid generator, a resin polymer, and a solvent. CAR is very sensitive to deep UV radiation and the exposed and unexposed regions differ greatly in their solubility in the developer solution.

13.1.5 Pattern Transfer

Figure 9 illustrates the steps to transfer IC patterns from a mask to a silicon wafer that has an insulating SiO_2 layer on its surface.⁸ The wafer is placed in a clean room that is typically illuminated with yellow light, since photoresists are not sensitive to wavelengths greater than $0.5 \mu\text{m}$. To ensure satisfactory adhesion of the resist, the surface must be changed from hydrophilic to hydrophobic. This change can be made by the application of an adhesion promoter, which can provide a chemically compatible surface for the resist. The most common adhesion promoter for silicon ICs is hexa-methylenedi-siloxane (HMDS). After the application of this adhesion layer, the wafer is held on a vacuum spindle and 2–3 cc of liquid resist is applied to the center of wafer. The wafer is then rapidly accelerated up to a constant rotational speed that is maintained for about 30 seconds. Spin speed is generally in the range of 1000–10,000 rpm (2000–5000 rpm is common) to coat a uniform film about 0.5 to $1 \mu\text{m}$ thick, as shown in Fig. 9a. The thickness of photoresist is correlated with its viscosity.

After the spinning step, the wafer is given a soft bake (typically at 90° – 120°C for 60–120 seconds) to remove the solvent from the photoresist film and to increase resist adhesion to the wafer. The wafer is aligned with respect to the mask in an optical lithographic system, and the resist is exposed to UV light, as shown in Fig. 9b. If a positive photoresist is used, the exposed resist is dissolved in the developer, as shown in the left side of Fig. 9c. Photoresist development is usually done by flooding the wafer with the developer solution. The wafer is then rinsed and dried. After development, postbaking at $\sim 100^\circ$ – 180°C may be required to increase the adhesion of the resist to the substrate. The wafer is then put in an ambient that etches the exposed insulation layer but does not attack the resist, as shown in Fig. 9d. Finally, the resist is stripped (e.g., using solvent or plasma oxidation), leaving behind an insulator image (or pattern) that is the same as the opaque image on the mask (left side of Fig. 9e).

For the negative photoresist, the procedures described are also applicable, except that the unexposed areas are removed. The final insulator image (right side of Fig. 9e) is the reverse of the opaque image on the mask.

The insulator image can be used as a mask for subsequent processing. For example, we use ion implantation to dope the exposed semiconductor region, but not the area covered by the insulator. The dopant pattern is a duplicate of the design pattern on the photomask (for a negative photoresist) or is its complementary pattern (for a positive photoresist). The complete circuit is fabricated by aligning the next mask in the sequence to the previous pattern and repeating the lithographic transfer process.

A related pattern-transfer process is the liftoff technique, shown in Fig. 10. A positive resist is used to form the resist pattern on the substrate (Fig. 10a and 10b). The film (e.g., aluminum) is deposited over the resist and the substrate (Fig. 10c); the film thickness must be smaller than that of the resist. Those portions of the film on the resist are removed by selectively dissolving the resist layer in an appropriate liquid etchant so that the overlying film is lifted off and removed (Fig. 10d). The liftoff technique is capable of high resolution and is used extensively for discrete devices such as high-power MESFETs. However, it is not as widely applicable for ultralarge-scale integration, in which dry etching is the preferred technique.

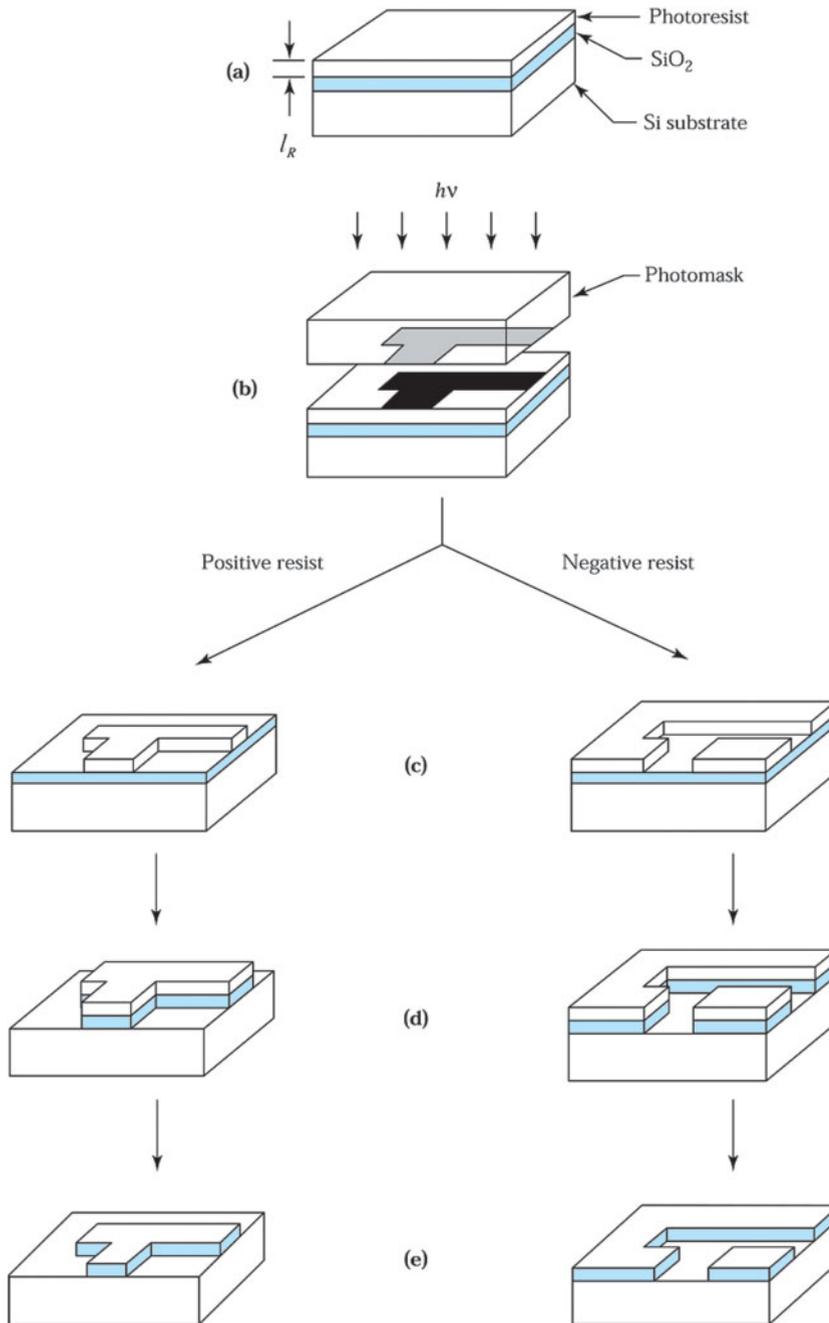


Fig. 9 Details of the optical lithographic pattern transfer process.⁸ (a) Application of resist. (b) Resist exposure through the mask. (c) Development of resist. (d) Etching of SiO₂. (e) Removal of resist.

Wet Photoresist Stripping

The photoresist can be stripped off with a strong acid such as H₂SO₄ or an acid-oxidant combination such as H₂SO₄-Cr₂O₃ attacking the resist but not the oxide or the Si. Other liquid strippers are organic-solvent strippers and alkaline strippers. Acetone can be used if the postbaking was not too long or at too high a temperature. At 120 °C we can

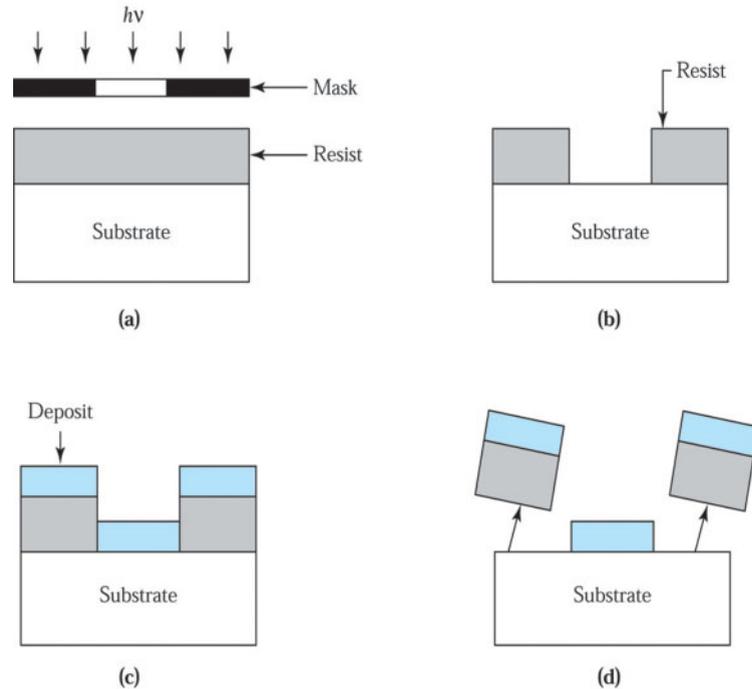


Fig. 10 Liftoff process for pattern transfer. (a) Resist exposure through the mask. (b) Resist. (c) Film deposition. (d) Liftoff.

use acetone. However, with a post-bake at 140 °C, the resist develops a tough skin and has to be burned away in oxygen plasma.

Dry Photoresist Stripping

Dry resist stripping (or ashing) can provide a cleaner surface than wet resist stripping. It also has fewer problems with toxic, flammable, and dangerous chemicals. The stripping rate is almost a constant and causes no undercutting and broadening of the resist. In addition, it is less corrosive with respect to metal features on the wafer.

There are three methods for dry resist stripping. Oxygen plasma stripping employs a low-pressure plasma discharge to split molecular oxygen (O_2) into its more reactive atomic form (O). This atomic oxygen converts an organic resist into a gaseous product that may be pumped away. In ozone strippers, ozone attacks the resist at atmosphere pressure. In UV/ozone stripping, the UV helps to break bonds in the resist, so that ozone can make a more efficient attack. Ozone strippers have the advantage that no plasma damage can occur on the devices in the process. The barrel plasma reactor has been used primarily for resist stripping and will be discussed in Section 13.5.

13.1.6 Resolution Enhancement Techniques

Optical lithography has been continuously challenged to provide better resolution, greater depth of focus (DOF), and wider exposure latitude in IC processing. These challenges have been met by reducing the wavelength of the exposure equipment and developing new resists. In addition, many resolution-enhancement techniques have been developed to extend the capability of optical lithography to even smaller feature lengths.

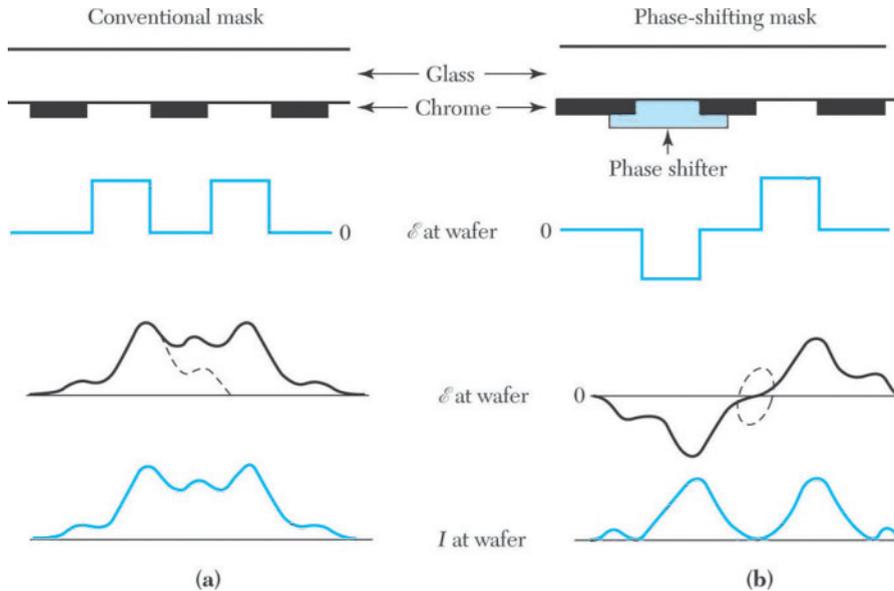


Fig. 11 The principle of phase-shift technology. (a) Conventional technology; (b) phase-shift technology.⁹

Phase Shift Technology

An important resolution enhancement technique is the phase-shifting mask (PSM). The basic concept is shown⁹ in Fig. 11. For a conventional mask, the electric field has the same phase at every aperture (clear area) in Fig. 11a. Diffraction and the limited resolution of the optical system spread the electric field at the wafer, as shown by the dotted lines. Interference between waves diffracted by the adjacent apertures enhances the field between them. Because the intensity I is proportional to the square of the electric field, it becomes difficult to separate the two images that are projected close to one another. The phase-shift layer that covers adjacent apertures reverses the sign of the electric field, as shown in Fig. 11b. Because the intensity at the mask is unchanged, the electric field of the images at the wafer can be cancelled. Therefore, images that are projected close to one another can be separated. A 180° phase change can be obtained by using a transparent layer with the thickness of $d = \lambda / 2(\bar{n} - 1)$, where \bar{n} is the refractive index and λ is the wavelength, that covers one aperture, as shown in Fig. 11b.

Optical Proximity Correction

High-performance optical projection imaging for lithography is strongly impacted by diffraction effects. The individual pattern features do not image independently, but rather interact with neighboring pattern features. The result from the diffraction overlap is the so-called proximity effect. The proximity effect becomes much more prominent as the feature sizes and spaces between the feature sizes approach the resolution limits of the projection optics.

A resolution-enhancement technique to minimize this effect is optical proximity correction (OPC), which uses modified shapes of adjacent subresolution geometry to compensate for image errors due to diffraction effects. For example, a line with a width near the resolution limit will print a line with round corners as shown as Fig. 12a due to the diffraction effect. Modifying the edge of the line pattern with additional geometrics at the corners as shown in Fig. 12b will help print a more accurate line. The addition of OPC features to the mask layout allows tighter design rules and significantly improves process reliability and yield.

Immersion Lithography

As mentioned in Sec. 13.1.2, immersion lithography is an advanced photolithography system in which the usual air gap between the lens and the wafer surface is replaced with a liquid medium that has a refractive index greater

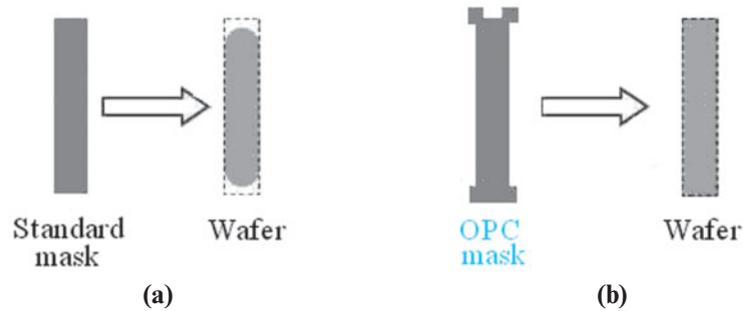


Fig. 12 Optical proximity effects. (a) Round corners by standard mask. (b) Accurate line shape by OPC mask.

than air. The resolution can be enhanced by increasing the numerical aperture (Eq. 2), which is proportional to the refractive index of the image medium (Eq. 3). Therefore, the resolution is increased by a factor equal to the refractive index. Current immersion-lithography equipment uses highly purified water ($\bar{n} = 1.33$) for this liquid to fabricate new-generation nano-scaled CMOS ICs. Immersion lithography is being developed for processes below 32 nm.

► 13.2 NEXT-GENERATION LITHOGRAPHIC METHODS

Why is optical lithography so widely used and what makes it such a promising method? The reasons are that it has high throughput, good resolution, low cost, and ease of operation. However, due to IC process requirements for features below 100 nm, optical lithography has some limitations not yet solved. Although we can use PSM, OPC, or immersion lithography to extend its useful span, the complexity of mask production and mask inspection can not be easily resolved. In addition, the cost of the masks is very high. Therefore, we need to find postoptical lithography to process nanometer ICs. Various types of next-generation lithographic methods for IC fabrication are discussed in this section.

13.2.1 Electron-Beam Lithography

Electron-beam lithography is primarily used to produce photomasks. Relatively little equipment is dedicated to direct exposure of the resist by a focused electron beam without a mask. Figure 13 shows a schematic of an electron-beam lithography system.¹⁰ The electron gun is a device that can generate a beam of electrons with a suitable current density. A tungsten thermionic-emission cathode or single-crystal lanthanum hexa-boride (LaB_6) is used for the electron gun. Condenser lenses are used to focus the electron beam to a spot size 10–25 nm in diameter. Beam-blanking plates that turn the electron beam on and off and beam deflection coils are computer controlled and operated at MHz or higher rates to direct the focused electron beam to any location in the scan field on the substrate. Because the scan field (typically 1 cm) is much smaller than the substrate diameter, a precision mechanical stage is used to position the substrate to be patterned.

The advantages of electron-beam lithography include the generation of nanometer resist geometries, highly automated and precisely controlled operation, greater depth of focus than available from optical lithography, and direct patterning on a semiconductor wafer without using a mask. The disadvantage is that electron beam lithographic machines have low throughput—approximately 2 wafers per hour at less than 100 nm resolution. This throughput is adequate for the production of photomasks, for situations that require small numbers of custom circuits, and for design verification. However, for maskless direct writing, the machine must have the highest possible throughput and therefore the largest beam diameter possible consistent with the minimum device dimensions.

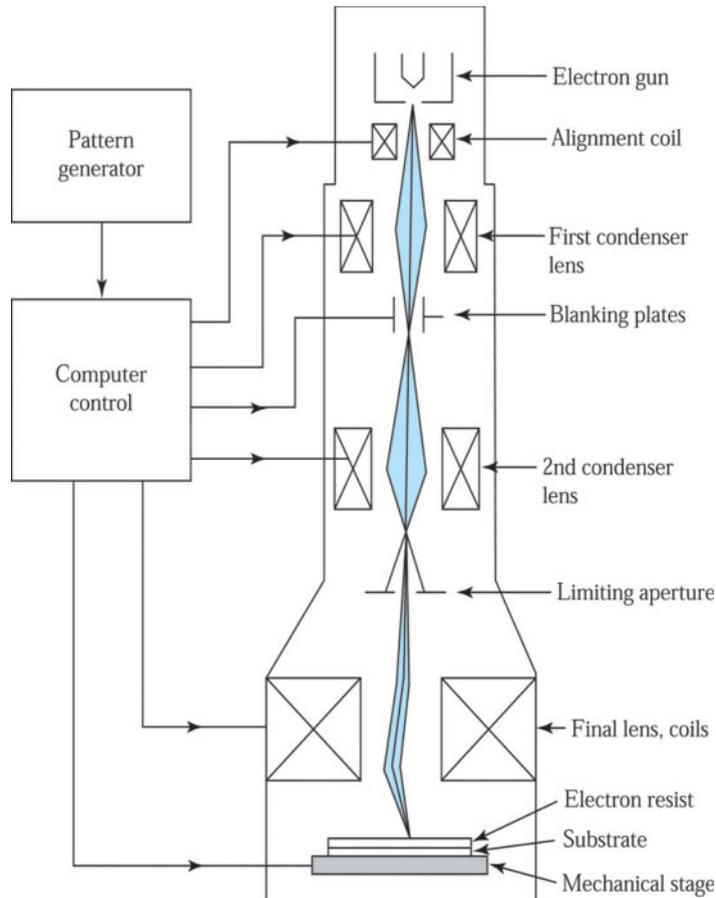


Fig. 13 Schematic of an electron-beam machine.¹⁰

There are basically two ways to scan the focused electron beam: raster scan and vector scan.¹¹ In a raster scan system, resist patterns are written by a beam that moves through a regular mode, vertically oriented, as shown in Fig. 14a. The beam scans sequentially over every possible location on the mask and is blanked (turned off) where no exposure is required. All patterns on the area to be written must be subdivided into individual addresses, and a given pattern must have a minimum incremental interval that is evenly divisible by the beam address size.

In the vector scan system, as shown in Fig. 14b, the beam is directed only to the requested pattern features and jumps from feature to feature, rather than scanning the whole chip, as in raster scan. For many chips, the average exposed region is only 20% of the chip area, so we can save time by using a vector-scan system.

Figure 14c shows several types of electron beams employed for e-beam lithography: the Gaussian spot beam (round beam), variable-shaped beam, and cell projection. In variable-shaped beam system, the patterning beam has a rectangular cross section of variable size. Therefore, the vector scan method using variable-shaped beam has higher throughput than the conventional Gaussian spot beam. It is also possible to pattern a complex geometric shape in one exposure with an electron beam system; this is called cell projection, as shown in the far right of Fig. 14c. The cell projection technique¹² is particularly suitable for highly repetitive designs, as in MOS memory cells, because several memory cell patterns can be exposed at once. Cell projection has not yet achieved the throughput of optical exposure equipment.

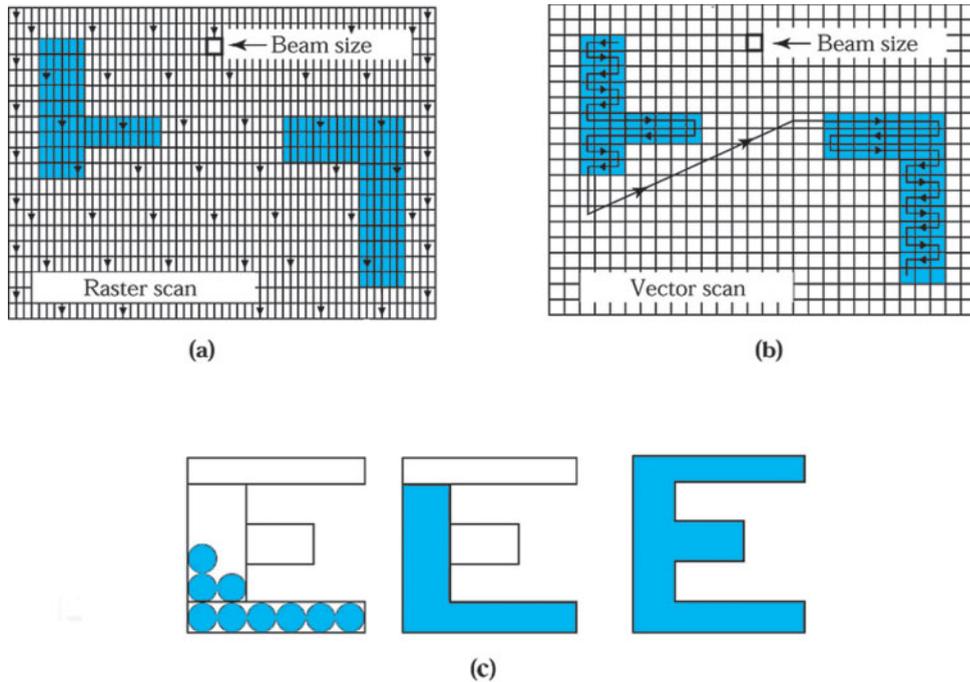


Fig. 14 (a) Raster-scan writing scheme; (b) vector-scan writing schemes; (c) shapes of electron beam: round, variable, cell projection.¹²

Electron Resist

Electron resists are polymers. The behavior of an electron-beam resist is similar to that of a photoresist: that is, a chemical or physical change is induced in the resist by irradiation. This change allows the resist to be patterned. For a positive electron resist, the polymer-electron interaction causes chemical bonds to be broken (chain scission) to form shorter molecular fragments, as shown¹³ in Fig. 15a. As a result, the molecular weight is reduced in the irradiated area, which can be dissolved subsequently in a developer solution that attacks the low-molecular-weight material. Common positive electron resists include poly-methyl methacrylate (PMMA) and poly-butene-1 sulfone (PBS). Positive electron resists can achieve resolutions of 0.1 μm or better.

For a negative electron resist, the irradiation causes radiation-induced polymer linking, as shown in Fig. 15b. The cross linking creates a complex three-dimensional structure with a molecular weight higher than that of the nonirradiated polymer. The nonirradiated resist can be dissolved in a developer solution that does not attack the high-molecular-weight material. Poly-glycidyl methacrylate-co-ethyl acrylate (COP) is a common negative electron resist. COP, like most negative photoresists, also swells during development, so the resolution is limited to about 1 μm .

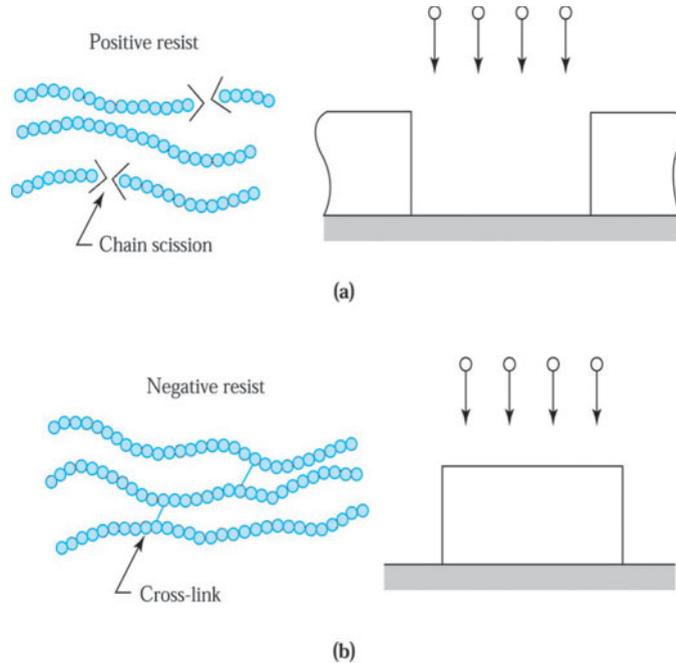


Fig. 15 Schematic of (a) positive and (b) negative resists used in electron-beam lithography.¹³

The Proximity Effect

In optical lithography, the resolution is limited by diffraction of light. In electron-beam lithography, the resolution is limited not by diffraction (because the wavelengths associated with electrons of a few keV and higher energies are less than 0.1 nm) but by electron scattering. When electrons penetrate the resist film and underlying substrate, they undergo collisions that lead to energy losses and path changes. Thus, the incident electrons spread out as they travel through the material until either all of their energy is lost or they leave the material because of backscattering.

Figure 16a shows computed electron trajectories of 100 electrons with initial energy of 20 keV incident at the origin of a 0.4 μm PMMA film on a thick silicon substrate.¹⁴ The electron beam is incident along the z -axis, and all trajectories have been projected onto the xz plane. This figure shows qualitatively that the electrons are distributed in an oblong pear-shaped volume with a diameter on the same order of magnitude as the electron penetration depth ($\sim 3.5 \mu\text{m}$). Also, there are many electrons that undergo backscattering collisions and travel backward from the silicon substrate into the PMMA resist film and leave the material.

Figure 16b shows the normalized distributions of the forward-scattering and backscattering electrons at the resist-substrate interface. Because of the backscattering, electrons can irradiate several micrometers away from the center of the exposure beam. Since the dose of a resist is given by the sum of the irradiations from all surrounding areas, the electron-beam irradiation at one location will affect the irradiation in neighboring locations. This phenomenon is called the *proximity effect*. The proximity effect places a limit on the minimum spacings between pattern features. To correct for the proximity effect, patterns are divided into smaller segments. The incident electron dose in each segment is adjusted so that the integrated dose from all its neighboring segments is the correct exposure dose. This approach further decreases the throughput of the electron-beam system because of the additional computer time required to expose the subdivided resist patterns.

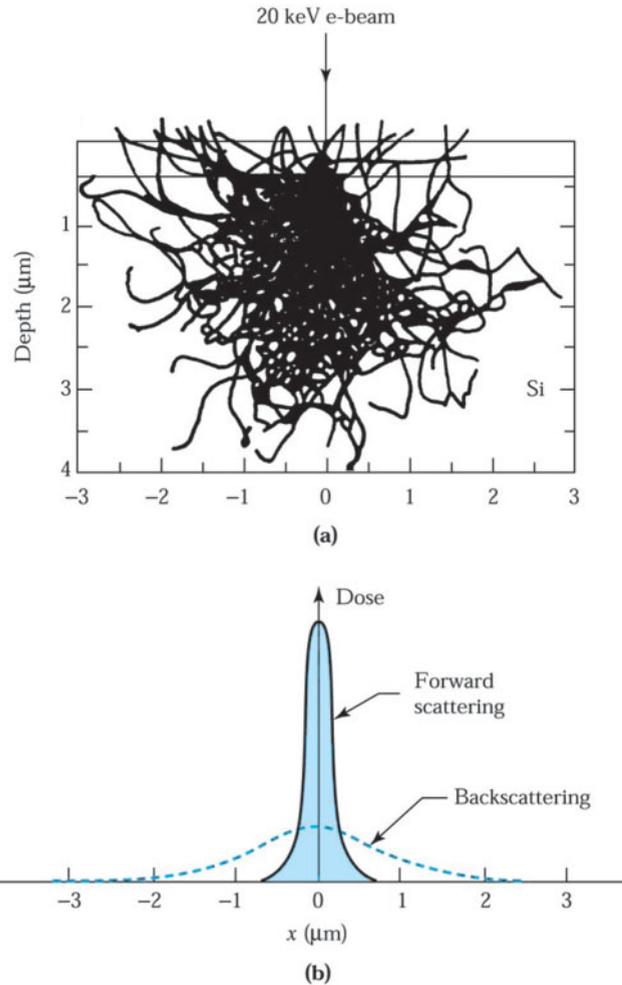


Fig. 16 (a) Simulated trajectories of 100 electrons in PMMA for a 20-keV electron beam.¹⁴ (b) Dose distribution for forward scattering and backscattering at the resist-substrate interface.

13.2.2 Extreme-Ultraviolet Lithography

Extreme-ultraviolet (EUV) lithography is a promising next-generation lithographic technology to extend minimum linewidths below 30 nm without throughput loss.¹⁵ Figure 17 shows a schematic diagram of an EUV lithographic system. A laser-produced plasma or synchrotron radiation can serve as the EUV source of $\lambda = 10\text{--}14$ nm EUV light. The EUV radiation is reflected by a mask that is produced by patterning an absorber material deposited on a multilayer-coated flat silicon or glass-plate mask blank. EUV radiation is reflected from the nonpatterned regions (i.e., nonabsorbing regions) of the mask through a $4\times$ reduction camera and imaged into a thin layer of resist on the wafer.

Since the EUV radiation beam is narrow, the mask must be scanned by the beam to illuminate the entire pattern field that describes the circuit mask layer. Also, for a $4\times$ four-mirror (i.e., the two-paraboloid, one-ellipsoid, and one-plane mirrors) reduction camera, the wafer must be scanned at one-fourth the mask speed in a direction opposite to the mask movement to reproduce the image field on all chip sites on the wafer surface. A precision system is required to perform

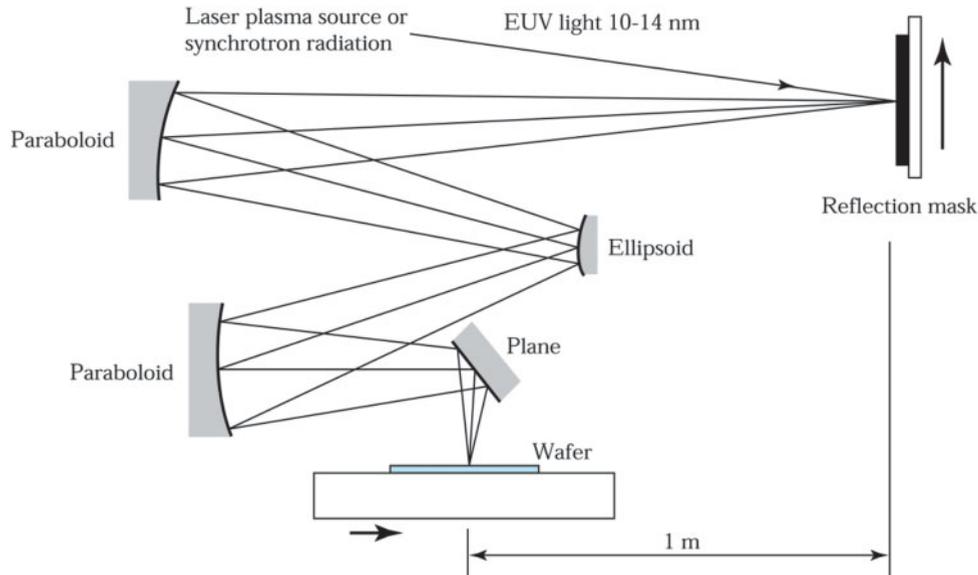


Fig. 17 Schematic representation of an extreme-ultraviolet (EUV) lithography system.¹⁵

the chip-site alignment and to control the wafer and mask stage movements and exposure dose during the scanning process. EUV lithography is capable of printing 50-nm features with PMMA resist using 13-nm radiation. However, the production of EUV exposure equipment has a number of challenges. Since EUV is strongly absorbed in all materials, the lithography process must be performed in vacuum. The camera must use reflective-lens elements, and the mirrors must be coated with multilayer coatings that produce distributed quarter-wave Bragg reflectors. In addition, the mask blank must also be multilayer coated to maximize its reflectivity at $\lambda = 10\text{--}14\text{ nm}$.

13.2.3 Ion-Beam Lithography

Ion-beam lithography can achieve higher resolution than optical or electron-beam lithographic techniques because ions have a greater mass and therefore scatter less than electrons. The most important application is the repair of masks for optical lithography, a task for which commercial systems are available.

The computer-simulated trajectories of 50 H^+ ions implanted at 60 keV into PMMA and various substrates¹⁶ shows that the spread of the ion beam at a depth of $0.4\text{ }\mu\text{m}$ is only $0.1\text{ }\mu\text{m}$ in all cases (compare with Fig. 16 for electrons). The backscattering is completely absent for the silicon substrate, and there is only a small amount of backscattering for the gold substrate. However, ion-beam lithography may suffer from a random (or stochastic) space-charge effect, causing broadening of the ion beam.

There are two types of ion-beam lithography systems: a scanning focused-beam system and a mask-beam system. The former system is similar to the electron-beam machine (Fig. 13), in which the ion source can be Ga^+ or H^+ . The latter system is similar to an optical $5\times$ reduction projection step-and-repeat system, which projects 100 keV light ions such as H_2^+ through a stencil mask.

13.2.4 Comparison of Various Lithographic Methods

The lithographic methods discussed above all have 100 nm or better resolution. However, each method has its own limitations. For IC fabrication, many mask levels are involved. However, it is not necessary to use the same lithographic method for all levels. A mix-and-match approach can take advantage of the unique features of each lithographic process to improve resolution and to maximize throughput. For example, a 4:1 EUV method can be used for the most critical mask levels, whereas 4:1 or 5:1 optical system can be used for the rest.

According to the Roadmap of the Semiconductor Industry Association, IC manufacturing technology will reach the 15 nm generation around 2020. With each new technology generation, lithography has become an even more important key driver for the semiconductor industry because of the requirements of smaller feature size and tighter overlay tolerance. In addition, lithography equipment costs have become higher relative to the total equipment costs for IC manufacturing facility. Currently, the technology development of next-generation lithography is conducted by multinational research projects or industrial partners.

▶ 13.3 WET CHEMICAL ETCHING

Wet chemical etching is used extensively in semiconductor processing. Starting from the sawed semiconductor wafers, chemical etchants are used for lapping and polishing to give an optically flat, damage-free surface. Prior to thermal oxidation or epitaxial growth, the semiconductor wafers are chemically cleaned to remove contamination that results from handling and storing. Wet chemical etchings are especially suitable for blanket etches (i.e., over the whole wafer surface) of polysilicon, oxide, nitride, metals, and III-V compounds.

The mechanisms for wet chemical etching involve three essential steps: the reactants are transported by diffusion to the reacting surface, chemical reactions occur at the surface, and the products from the surface are removed by diffusion. Both agitation and the temperature of the etchant solution will influence the etch rate, which is the amount of film removed by etching per unit time. In IC processing, most wet chemical etchings proceed by immersing the wafers in a chemical solution or by spraying the wafers with the etchant solution. For immersion etching, the wafer is immersed in the etch solution. Mechanical agitation is usually required to ensure etch uniformity and a consistent etch rate. Spray etching has gradually replaced immersion etching because it greatly increases the etch rate and uniformity by constantly supplying fresh etchant to the wafer surface.

Etch rates must be uniform across a wafer, from wafer to wafer, from run to run, and for any variations in feature sizes and pattern densities. Etch rate uniformity is given by:

$$\text{Etch rate uniformity (\%)} = \frac{(\text{maximum etch rate} - \text{minimum etch rate})}{\text{maximum etch rate} + \text{minimum etch rate}} \times 100\%. \quad (8)$$

▶ EXAMPLE 3

Calculate the Al average etch rate and etch rate uniformity on a 300 mm diameter silicon wafer, assuming the etch rates at the center, left, right, top, and bottom of the wafer are 750, 812, 765, 743, and 798 nm/min, respectively.

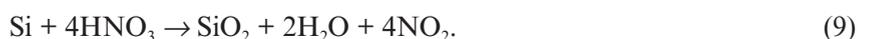
SOLUTION

$$\text{Al average etch rate} = (750 + 812 + 765 + 743 + 798) \div 5 = 773.6 \text{ nm/min.}$$

$$\text{Etch rate uniformity} = (812 - 743) \div (812 + 743) \times 100\% = 4.4 \%. \quad \blacktriangleleft$$

13.3.1 Silicon Etching

For semiconductor materials, wet chemical etching usually starts with oxidation followed by dissolution of the oxide by a chemical reaction. For silicon, the most commonly used etchants are mixtures of nitric acid (HNO_3) and hydrofluoric acid (HF) in water or acetic acid (CH_3COOH). Nitric acid oxidizes silicon to form a SiO_2 layer.¹⁷ The oxidation reaction is



Hydrofluoric acid is used to dissolve the SiO_2 layer. The reaction is:



Water can be used as a diluent for this etchant. However, acetic acid is preferred because it reduces the dissolution of the nitric acid.

Some etchants dissolve a given crystal plane of single-crystal silicon much faster than another plane; this results in orientation-dependent etching.¹⁸ For a silicon lattice, the (111)-plane has more available bonds per unit area than the (110)- and (100)-planes; therefore, the etch rate is expected to be slower for the (111)-plane. A commonly used orientation-dependent etch for silicon consists of a mixture of KOH in water and isopropyl alcohol. For example, a solution with 19 wt% KOH in deionized (DI) water at about 80°C removes the (100)-plane at a much greater rate than the (110)- and (111)-planes. The ratio of the etch rates for the (100)-, (110)-, and (111)-planes is 100:16:1.

Orientation-dependent etching of <100>-oriented silicon through a patterned silicon dioxide mask creates precise V-shaped grooves,¹⁰ the edges being (111)-planes at an angle of 54.7° from the (111)-surface, as shown at the left of Fig. 18a. If the window in the mask is sufficiently large or if the etching time is short, a U-shaped groove will be formed, as shown at the right of Fig. 18a. The width of the bottom surface is given by

$$W_b = W_0 - 2l \cot 54.7^\circ$$

or

$$\boxed{W_b = W_0 - \sqrt{2} l}, \quad (11)$$

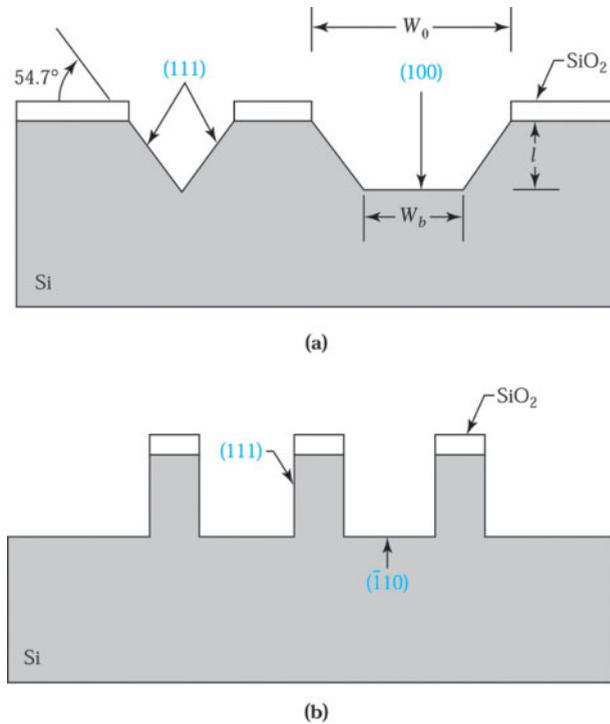


Fig. 18 Orientation-dependent etching. (a) Through window patterns on <100>-oriented silicon; (b) through window patterns on <100>-oriented silicon.¹⁸

where W_0 is the width of the window on the wafer surface and l is the etched depth. If $\langle 110 \rangle$ -oriented silicon is used, essentially straight-walled grooves with sides of (111)-planes can be formed, as shown in Fig. 18b. We can use the large orientation dependence in the etch rates to fabricate device structures with submicron feature lengths.

13.3.2 Silicon Dioxide Etching

The wet etching of silicon dioxide is commonly achieved in a dilute solution of HF with or without the addition of ammonium fluoride (NH_4F). Adding NH_4F is referred to as a buffered HF solution (BHF), also called buffered-oxide-etch (BOE). The addition of NH_4F to HF controls the pH value and replenishes the depleted fluoride ions, thus maintaining stable etching performance. The overall reaction for SiO_2 etching is the same as that in Eq. 10. The etch rate of SiO_2 etching depends on the etchant solution, etchant concentration, agitation, and temperature. In addition, density, porosity, microstructure, and the presence of impurities in the oxide also influence the etch rate. For example, a high concentration of phosphorus in the oxide results in a rapid increase in the etch rate, and a loosely structured chemical-vapor deposition (CVD) or sputtered oxide exhibits a faster etch rate than thermally grown oxide.

Silicon dioxide can also be etched in vapor-phase HF. Vapor-phase-HF oxide-etch technology has a potential for etching feature lengths below 100 nm because the process can be well controlled.

13.3.3 Silicon Nitride and Polysilicon Etching

Silicon nitride films can be etched at room temperature in concentrated HF or buffered HF and in a boiling H_3PO_4 solution. Selective etching of nitride to oxide is done with 85% H_3PO_4 at 180°C because this solution attacks silicon dioxide very slowly. The etch rate is typically 10 nm/min for silicon nitride, but less than 1 nm/min for silicon dioxide. However, photoresist adhesion problems are encountered when etching nitride with boiling H_3PO_4 solution. Better patterning can be achieved by depositing a thin oxide layer on top of the nitride film before resist coating. The resist pattern is transferred to the oxide layer, which then acts as a mask for subsequent nitride etching.

Etching polysilicon is similar to etching single-crystal silicon. However, the etch rate is considerably larger because of grain boundaries. The etch solution is usually modified to ensure that it does not attack the underlying gate oxide. Dopant concentrations and temperature may affect the etch rate of polysilicon.

13.3.4 Aluminum Etching

Aluminum and aluminum alloy films are generally etched in heated solutions of phosphoric acid, nitric acid, acetic acid, and DI water. The typical etchant is a solution of 73% H_3PO_4 , 4% HNO_3 , 3.5% CH_3COOH , and 19.5% DI water at 30°–80°C. The wet etching of aluminum proceeds as follows: HNO_3 oxidizes aluminum, and H_3PO_4 then dissolves the oxidized aluminum. The etch rate depends on etchant concentration, temperature, agitation of the wafers, and impurities or alloys in the aluminum film. For example, the etch rate is reduced when copper is added to the aluminum.

Wet etching of insulating and metal films is usually done with the similar chemicals that dissolve these materials in bulk form. Generally, film materials will be etched more rapidly than their bulk counterparts. Also, the etch rates are higher for films that have a poor microstructure, built-in stress, departure from stoichiometry, or have been irradiated. Some useful etchants for insulating and metal films are listed in Table 1.

13.3.5 Gallium Arsenide Etching

A wide variety of etches has been investigated for gallium arsenide; however, few of them are truly isotropic.¹⁹ This is because the surface activities of the (111)-Ga and (111)-As faces are very different. Most etches give a polished surface on the arsenic face, but the gallium face tends to show crystallographic defects and etches more slowly. The most commonly used etchants are the H_2SO_4 - H_2O_2 - H_2O and H_3PO_4 - H_2O_2 - H_2O systems. For an etchant with an 8:1:1 volume ratio of H_2SO_4 : H_2O_2 : H_2O , the etch rate is 0.8 $\mu\text{m}/\text{min}$ for the $\langle 111 \rangle$ -Ga face and 1.5 $\mu\text{m}/\text{min}$ for all other faces. For an etchant with 3:1:50 volume ratio of H_3PO_4 : H_2O_2 : H_2O , the etch rate is 0.4 $\mu\text{m}/\text{min}$ for $\langle 111 \rangle$ -Ga face and 0.8 $\mu\text{m}/\text{min}$ for all other faces.

TABLE 1 ETCHANTS FOR INSULATORS AND CONDUCTORS

| Material | Etchant composition | Etch rate (nm/min) |
|--------------------------------|--|--------------------|
| SiO ₂ | 28 ml of HF 170 ml of HF 113 g of NH ₄ F | Buffered HF 100 |
| | 15 ml of HF 10 ml of HNO ₃ 300 ml of H ₂ O | |
| Si ₃ N ₄ | Buffered HF | 0.5 |
| Al | H ₃ PO ₄ | 10 |
| | 4 ml of HNO ₃ 3.5 ml of CH ₃ COOH 73 ml of H ₃ PO ₄ 19.5 ml of H ₂ O | 30 |
| Au | 4 g KI | 1000 |
| Mo | 1 g of I ₂ 40 ml of H ₂ O 5 ml of H ₃ PO ₄ 2 ml of HNO ₃ 4 ml of CH ₃ COOH 150 ml of H ₂ O | 500 |
| | 1 ml of HNO ₃ 7 ml of HCl 8 ml of H ₂ O | |
| Pt | 34 g of KH ₂ PO ₄ 13.4 g of KOH 33 g of K ₃ Fe(CN) ₆ H ₂ O to make 1 liter | 160 |

► 13.4 DRY ETCHING

In pattern-transfer operations, a resist pattern is defined by a lithographic process to serve as a mask for etching of its underlying layer (Fig. 19a).²⁰ Most of the layer materials (e.g., SiO₂, Si₃N₄, and deposited metals) are amorphous or polycrystalline thin films. If they are etched in a wet chemical etchant, the etch rate is generally isotropic (i.e., the lateral and vertical etch rates are the same), as illustrated in Fig. 19b. If h_f is the thickness of the layer material and l the lateral distance etched underneath the resist mask, we can define the degree of anisotropy A_f by

$$A_f \equiv 1 - \frac{l}{h_f} = 1 - \frac{R_l t}{R_v t} = 1 - \frac{R_l}{R_v}, \quad (12)$$

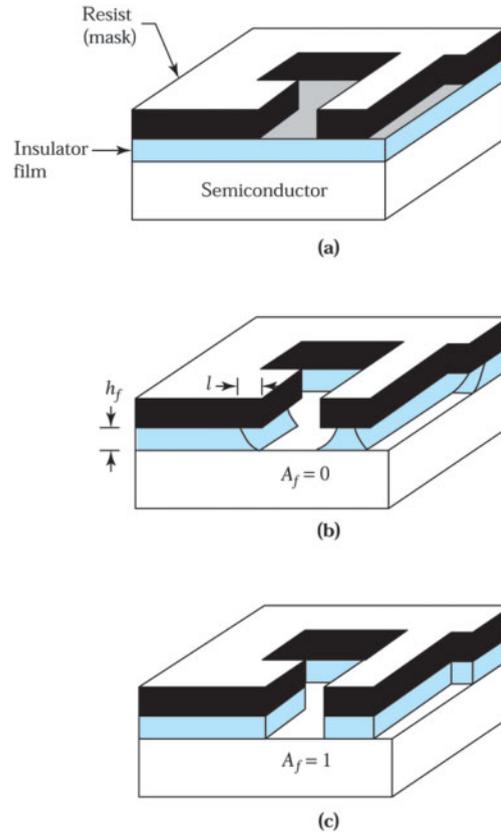


Fig. 19 (a) Resist pattern formation. Comparison of (b) wet chemical etching and (c) dry etching for pattern transfer.²⁰

where t is the time and R_l and R_v are the lateral and vertical etch rates, respectively. For isotropic etching, $R_l = R_v$ and $A_f = 0$.

The major disadvantage of wet chemical etching for pattern transfer is the undercutting of the layer underneath the mask, resulting in a loss of resolution in the etched pattern. In practice, for isotropic etching, the film thickness should be about one-third or less of the resolution required. If patterns are required with resolutions much smaller than the film thickness, anisotropic etching (i.e., $1 \geq A_f > 0$) must be used. In practice, the value of A_f is chosen to be close to unity. Figure 19c shows the limiting case where $A_f = 1$, corresponding to $l = 0$ (or $R_l = 0$).

To achieve high-fidelity transfer of the resist patterns required for ultralarge-scale integration processing ($A_f = 1$), dry etching methods have been developed. Dry etching is synonymous with plasma-assisted etching, which denotes several techniques that use plasma in the form of low-pressure discharges. Dry-etch methods include plasma etching, reactive ion etching (RIE), sputter etching, magnetically enhanced RIE (MERIE), reactive ion beam etching, and high-density plasma (HDP) etching.

13.4.1 Plasma Fundamentals

Plasma is a fully or partially ionized gas composed of equal numbers of positive and negative charges and a different number of unionized molecules. A simple capacitively coupled radio frequency (rf) plasma etcher schematically shown in Fig. 20 is used to demonstrate the plasma fundamentals. The cathode is capacitively coupled to an rf generator and the anode is grounded, similarly to the sputtering discussed in the previous chapter. The rf frequency is typically 13.56 MHz because of its non-interference with radio-transmitted signals. The plasma is initiated by free electrons always present in a gas, generated by cosmic rays, thermal excitation, or other means. The free electrons oscillate and gain kinetic energy from the rf electric field and collide with gas molecules. The energy transferred

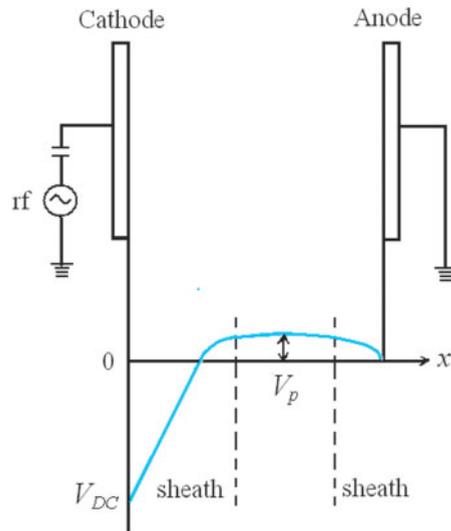


Fig. 20 Schematic system and approximate time-averaged potential distribution of a capacitively coupled rf plasma system.

in the collision causes the gas molecules to be ionized. When the applied voltage is larger than the breakdown potential of the gas, sustainable plasma is generated throughout the reaction chamber. The ionization rate is on the order of 10^{-4} to 10^{-6} .

Sheath

The formation of sheath in dry etching is similar to that in rf sputtering discussed in Chapter 12. Plasma is electrically neutral in the central region of the chamber. Electrons are more mobile than positive ions, and therefore more electrons are attracted to the front surface of the electrodes during the positive half cycle than positive ions in the negative half cycle. Therefore, the current is larger in the positive cycle than that in the negative cycle. The resultant electron current charges up the capacitively coupled electrode (powered electrode) since no charge can be transferred through the capacitor. The powered electrode (cathode) acquires an increasing negative bias voltage during successive cycles until the negative average voltage V_{DC} (also called 'self-bias') is sufficiently high to retard the electrons and the net charge arriving at the surface is zero. The magnitude of the powered electrode self-bias voltage depends on the amplitude and frequency of the voltage applied to the electrode. Since the powered electrode develops a negative self-bias, the plasma forms a compensating positive potential V_p relative to the grounded anode, as shown in the lower part of Fig. 20.

The voltage gradients near the cathode and anode form intense electric fields near plasma-electrode interfaces known as sheaths (also called dark space because the high-energy electrons there are more likely to cause ionization than light-generating excitation) that play a significant role in the plasma etching processes. Since typical sheaths are thin ($\sim 10 \mu\text{m}$ to 1 mm) and conformal with the electrode surface, positive ion energy gain is primarily in the direction normal to the surface and the ion beam is essentially unidirectional there. Anisotropic etching relies on bombardment of unidirectional energetic ions at the substrate surface, which can be placed on the cathode or anode. This is achieved in plasma etching reactors by accelerating positive ions in the sheath above the substrate surface. The asymmetric voltage distribution at the cathode and anode causes a very large field in front of the cathode in comparison with the field in front of the anode or in the glow region. Anisotropic etching is very strong on the cathode surface due to the very strong field there, and is weaker on anode surface due to the relatively weaker field.

13.4.2 Surface Chemistry

The plasma used for etching is not in thermal equilibrium. As a result, the temperature of electrons, which are the lightest component of the plasma, is substantially higher than the neutral gas and ion temperature. The electron temperature is approximately in the range of 20,000~100,000 K; the ion temperature might be up to 2,000 K, while neutral radicals and molecules are less than 1,000 K. These energetic electrons can therefore generate reactive radicals and ions and enhance chemical reactions that cannot be achieved by other means. Radicals produced during dissociation tend to be more reactive than the parent gases and these radicals can further enhance the surface processes and plasma chemistry.

Plasma etching has to satisfy many stringent requirements simultaneously, including control of feature sidewall and bottom surface profiles, etch selectivity to other exposed materials, uniformity of the etch process over large substrate surfaces, and interaction with preceding and following processing steps. The crucial points related to fundamental surface processes are physical sputtering, reactive ion etching (RIE), chemical etching, and polymer deposition.

Physical Sputtering

One of the simplest material removal processes is physical sputtering, which involves the bombardment of target material by energetic ions or neutrals. However, sputtering tends to be non-selective.

Reactive Ion Etching

Most plasma etching processes rely primarily on reactive ion etching for material removal. RIE involves simultaneous bombardment of energetic ions and reactive neutral radicals onto the material surface. Ions bombard the substrate surface almost normally and etching by the reactive neutral radicals occurs anisotropically. RIE is similar to sputtering, but more selective than physical sputtering due to its partially chemical nature from reactive neutral radicals.

Chemical Etching

A simple example of chemical plasma etching is Si etching using F, which has a high etch rate even at room temperature:



Chemical etching is often isotropic as incoming neutral etchants have a uniform angular distribution. However, for some crystalline materials, chemical etching can be sensitive to crystallographic orientation. During the fabrication of submicron-sized features in CMOS devices, chemical etching often cannot be tolerated due to its isotropic nature. Processing conditions are therefore chosen so as to minimize chemical etching.

Polymer Deposition

To generate small features, anisotropic etching also requires that etching take place only in the vertical direction with no etching in the horizontal direction. Although careful design of the plasma etching reactor and appropriate choice of etching gases can help to achieve these goals, one surface mechanism that has proven indispensable is polymer deposition. Presence of these films on vertical surfaces limits contact of the material surface with the etchant species to inhibit horizontal etching.

There are at least two mechanisms that can account for this buildup of sidewall passivation. The first is the deposition of polymeric material that is known to occur in plasma discharges with carbon-containing source gases. In the case of fluorine-containing Freons as source gases, this polymer deposition is linked to the formation of unsaturated CF_2 radicals generated by the plasma. The second source of material on feature sidewalls is the etch product species generated at horizontal surfaces exposed to ion bombardment. These products are frequently nonvolatile and can stick to and react with vertical surfaces not exposed to ion bombardment. This source of sidewall building is termed redeposition.

A series of sequential cross sections depicting the anisotropic etching of a feature with sidewall redeposition is shown in Fig. 21, in which the six sequential profiles result from five etch-redeposition-etch steps.



Fig. 21 The sequential formation, from left to right, of an etching feature profile in the presence of redeposition. The etching of horizontal surfaces and redeposition onto vertical surfaces are assumed to occur sequentially.

Substrate Temperature

Many of the above-mentioned fundamental surface processes take place simultaneously in etching processes. Plasma operating conditions must be carefully monitored to enhance or reduce the contribution of individual surface process and control the final results. One parameter that is particularly useful is the substrate temperature because many fundamental surface processes exhibit strong temperature dependence. For example, the chemical etching rate generally increases with surface temperature. Therefore, for processes in which both physical and chemical etching components are present, one can vary the degree of anisotropic or isotropic etching by varying the substrate temperature to control the feature profile.

13.4.3 Capacitively Coupled Plasmas Etchers

Dry etching technology in the IC industry has changed dramatically since the first application of plasma processing to photoresist stripping. A reactor for dry etching contains a vacuum chamber, pump system, power supply generators, pressure sensors, gas flow control units, and end-point detector. Each reactor uses a particular combination of pressure, electrode configuration and type, and source frequency to control the two primary etch mechanisms—chemical and physical. Higher etch rates and automation are required for most etchers used in IC fabrication. There are basically two groups of dry etchers based on how the plasma is produced: the capacitively coupled etchers and the inductively coupled etchers.¹

In the simplest form of a capacitively coupled plasma etcher, etchant gases are injected between two parallel metallic electrodes with symmetrical size and position to which voltage is applied on one electrode. The potential drop across the gas breaks it down and generates the plasma. A significant fraction of the input power is consumed by ions accelerating in the sheaths, and is dissipated at the electrode surfaces (or substrates placed on them) during ion bombardment. Therefore, a small fraction of the input power is used for plasma generation. The gas dissociation fraction is low and electron density is also low ($\sim 10^9$ to 10^{10} cm^{-3}). In addition, simple commercial capacitively coupled plasma etchers are typically operated at moderate gas pressures (~ 50 to 500 mTorr) and the scattering of gas prevents their use for fabrication of extremely small features.

As shown in Fig. 20, a wafer can be placed on the grounded electrode. This is the plasma etch mode with energetic ion bombardment since the plasma potential is always above the grounded potential. If a wafer is placed on the powered electrode (cathode), it is operated in the reactive ion etch mode with higher energetic ion bombardment due to higher self-bias V_{DC} . Physical and chemical etch mechanisms occur in both the plasma etch mode and the reactive ion etch mode. However, energies of bombarding ions are about ten times higher in the reactive ion etch mode.

Reactive Ion Etcher

A capacitively coupled plasma etcher operated in the reactive ion etch mode is called a reactive ion etcher (RIE) or a reactive sputter etcher (RSE). RIE has been extensively used in the microelectronic industry. The wafers are placed on the powered electrode (cathode). This allows the grounded electrode to have a significantly larger area, as shown in Fig. 22, and leads to significantly higher plasma-sheath potentials (20-500 V) at the wafer surface. The process can be explained as follows.

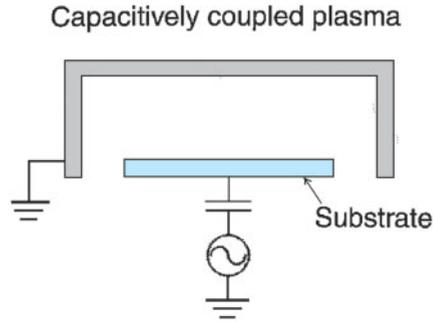


Fig. 22 Capacitively coupled plasma etcher with larger area for the grounded electrode.

The glow region of the plasma is a good electrical conductor. The dark spaces in plasma are areas of limited conductivity and can be modeled as capacitances, i.e., $C = A/d$, where A is the electrode area and d is the sheath thickness of the dark space. A voltage will split between two capacitances in series, i.e.

$$V_C/V_A = C_A/C_C = (A_A/d_A)/(A_C/d_C) \quad (14)$$

where V_C (C_C) and V_A (C_A) are the voltage drops (capacitances) over the sheath thicknesses of the dark space on the cathode and anode and A_C , A_A are the areas of the cathode and anode. The current between two electrodes is dominated by space-charge-limited current in a capacitively coupled plasma system. The space-charge-limited current (described in Section 2.7, Chapter 2) of positive ions must be equal on both anode and cathode, i.e.

$$V_C^{3/2}/d_C^2 = V_A^{3/2}/d_A^2 \quad (15)$$

Therefore,

$$V_C/V_A = (A_A/A_C)^4 \quad (16)$$

That is to say, the potential difference across the dark space of each electrode will be the same if the electrodes are of similar area. The increase of the relative surface area of the grounded electrode can increase the sheath voltage at the powered electrode. The etching rate can be much enhanced, but the etch selectivity of this system is relatively low because of strong physical sputtering. However, selectivity can be improved by choosing the proper etch chemistry, for example by polymerizing the silicon surface with fluorocarbon polymers to obtain selectivity of SiO_2 over silicon.

Magnetically Enhanced Reactive Ion Etcher

In the magnetically enhanced reactive ion etcher (MERIE), the magnetic field crossed with the electric field reduces electron mobility towards the electrodes and their loss there. Densities of electrons and other species in the plasma are therefore larger in MERIE reactors for the same input power, which enhances the material etch rate. For a given power, higher electron (and ion) densities in MERIE reactors will consume more fractional power and therefore smaller fractional power will be used to accelerate ions in the sheaths. Consequently, ion-bombardment-induced damage on the substrate and electrode surfaces diminishes. Etch uniformity is improved in MERIE reactors by either shaping the applied magnetic field or rotating it physically or electrically. Magnetically enhanced reactive ion etchers have been used extensively for dielectric etching in the semiconductor industry.

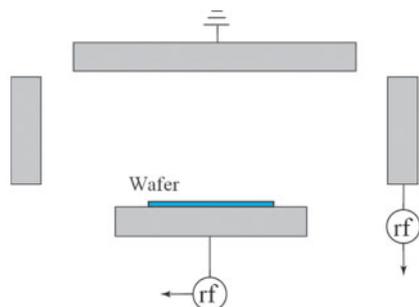


Fig. 23 Schematic of a triode reactive ion etch reactor with two different radio-frequency power sources.

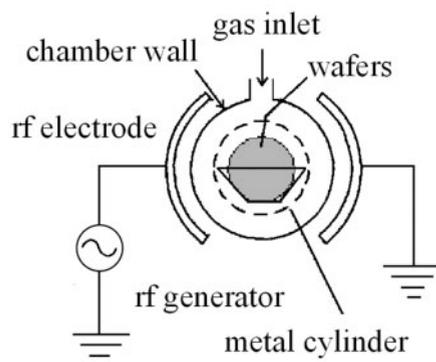


Fig. 24 Schematic of a typical barrel reactor.

Triode Reactive Ion Etcher

Another innovation in capacitively coupled plasma etcher design is the use of two (or more) sources at different frequencies as shown in Fig. 23. At the same input power, capacitively coupled plasmas are generated more efficiently at higher frequencies due to higher collision frequencies, and higher electron densities are accumulated at lower frequencies and hence higher self-biases are induced at the cathode. A high-frequency source (25 MHz and above) is therefore used in dual-frequency plasma systems to generate the plasma efficiently, while a low-frequency source (typically a few MHz or lower) accelerates ions. One can therefore obtain a higher plasma density relative to a simpler capacitive plasma system, and independently control ion energy as well.

Barrel Plasma Etcher

The barrel plasma reactor has been used primarily for resist stripping, as discussed in Ch. 12. It is one of the earliest plasma systems. The barrel reactor has a cylindrical design operated at a pressure of about 0.1 to 1 Torr. The power is applied on electrodes placed on both sides of the cylinder. An inner metal cylinder with holes can confine the plasma to the region between the metal cylinder and chamber wall (Fig. 24). The etchant species in the plasma diffuse through holes to etching area, while the energetic ions and electrons of the plasma cannot enter this region. Wafers are placed vertically on a quartz boat with a small separation between wafers, and placed parallel to the electric field to minimize physical etching. The etching is almost purely chemical with isotropic etching and high selectivity.

13.4.4 Inductively Coupled Plasma Etchers

Inductively coupled plasma (ICP) etchers were developed in the early 1990s to address the difficult process requirements of high-aspect-ratio (AR) oxide etch with high selectivity. ICP etchers are operated at lower gas pressure (~3 to 50 mtorr) than capacitively coupled plasma etchers. The lower pressure reduces gas collisions that cause loss of the etching profile. It also increases the mean free path of etchants and etching byproducts, and they can move easily into and out of high-aspect ratio features. However, the lower pressure also reduces the etch rate due

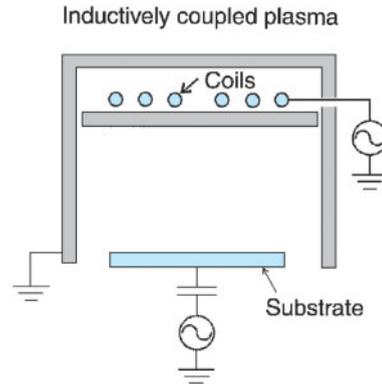


Fig. 25 Inductively coupled plasma etcher.

to the reduction in ion density. Therefore, high-density plasma (HDP) is needed to generate sufficient active species for an acceptable etch rate at lower pressures. The gas dissociation rate of HDP can reach about 10% compared with 0.1% of capacitively coupled plasmas etchers.

ICP etchers utilize a set of coils that are physically separated from the gas through a dielectric window as shown in Fig. 25. Radio-frequency current through the coils generates an electromagnetic wave that penetrates the plasma chamber, azimuthally accelerates electrons, and generates the plasma. As most of the input power is consumed by electrons, the electron density is substantially larger ($\sim 10^{11}$ - 10^{12} cm^{-3}) in ICP etchers than in capacitively coupled plasmas. Therefore, the ICP etcher is a high-density plasma (HDP) etcher.

In addition, a second source can be used in ICP etchers to separately bias the substrate during etching and impart energy to bombarding ions to enhance the etch rate. As separate sources are used for plasma generation and ion acceleration, high-aspect-ratio oxide etching is possible. Although a higher degree of dissociation enhances etch rate, it has detrimental effects on material selectivity in many cases.

Electron Cyclotron Resonance (ECR) Plasma Etcher

The ECR plasma etcher shown in Fig. 26, which is similar to an ICP etcher, uses resonant wave-plasma interaction. In an ECR etcher, microwave (typically at 2.45 GHz) is launched into a magnetized chamber containing the etchant gas at low pressure (<10 mTorr). Electron cyclotron resonance occurs at spatial locations where the local electron cyclotron frequency (eB/m_e) matches the applied frequency. By carefully designing the magnetic field profile, one can obtain high-density uniform plasma above the substrate surface. Plasma densities in ECR etchers are higher or comparable to ICP reactors. The ECR etcher is also a HDP etcher. Electron cyclotron resonance etchers are also operated at lower gas pressures than capacitively coupled plasma etchers and allow independent biasing of the substrate. Similarly to ICP, ECR etchers are characterized by high degrees of gas dissociation.

Neutral Beam Plasma Etcher

Because of the presence of charged species or ultraviolet radiation in plasma, electrical damage to circuits on the substrate remains a constant concern during plasma etching. To alleviate this problem, plasma etch sources that rely on energetic neutral beams have been developed in recent years. A typical neutral beam source is shown in Fig. 27, and similar designs have been used in the past for ion milling applications. Plasma of the appropriate gas is generated through conventional means and is then allowed to seep through holes in the electrodes. Ions can then be accelerated and neutralized before they bombard the substrate, resulting in energetic neutral-species bombardment on the substrate. Neutral beam sources are under development currently and have not yet been used for high-volume production. Another technique that has been utilized to alleviate plasma charge damage is

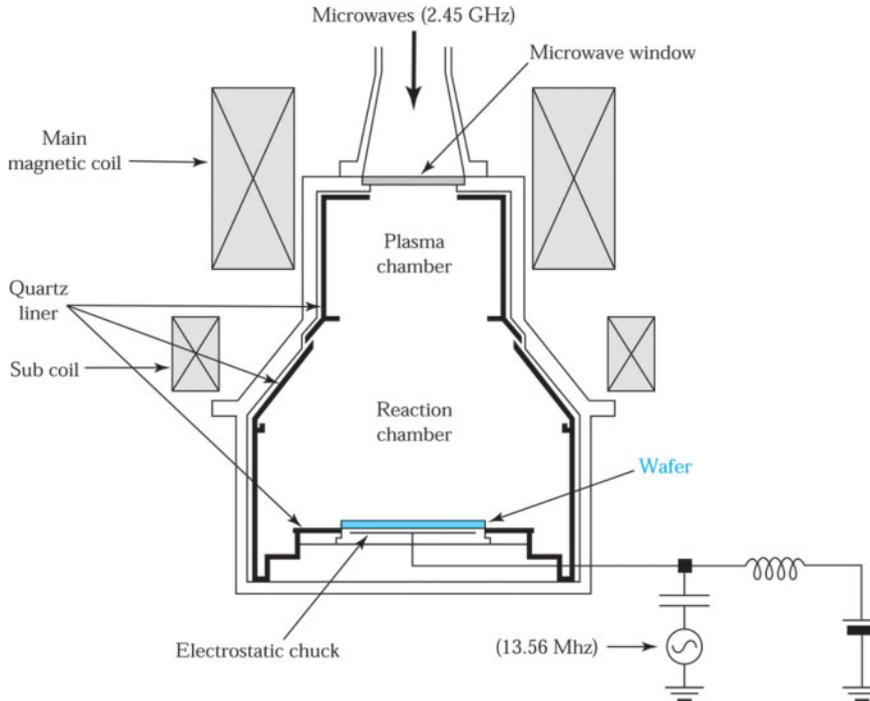


Fig. 26 Schematic of an electron cyclotron resonance reactor.²¹

to generate the plasma remotely, i.e., away from the substrate, and transport neutral species to the substrate so that ions are either excluded or neutralized. Remote plasma sources or chemical downstream etchers are used for many plasma cleaning and material treatment applications. These etchers are also useful for high-rate removal of blanket films that do not have any patterned features. Their application for anisotropic etching applications is, however, limited due to the broad angular distribution of neutral etchants.

Single-Wafer Etcher

For modern circuits with nanometer feature sizes, etching processes are more critical. More vertical profiles, better linewidth control, higher selectivity, and better uniformity are necessary. One approach to this problem is to use single-wafer etchers that etch one wafer at a time. Single-wafer etchers can tailor the electrode geometry and gas flow to maximize etch uniformity across the wafer. These machines are easily automated to perform wafer cassette-to-cassette operations so that no operator handling is required. They can incorporate a

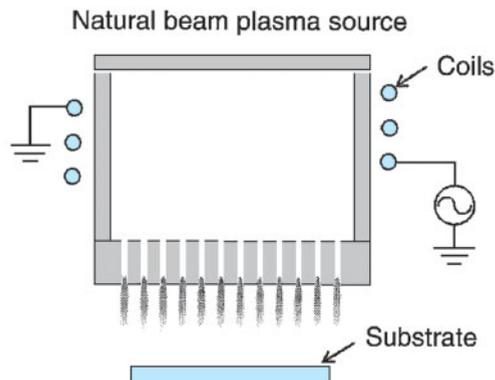


Fig. 27 Neutral beam plasma etcher.

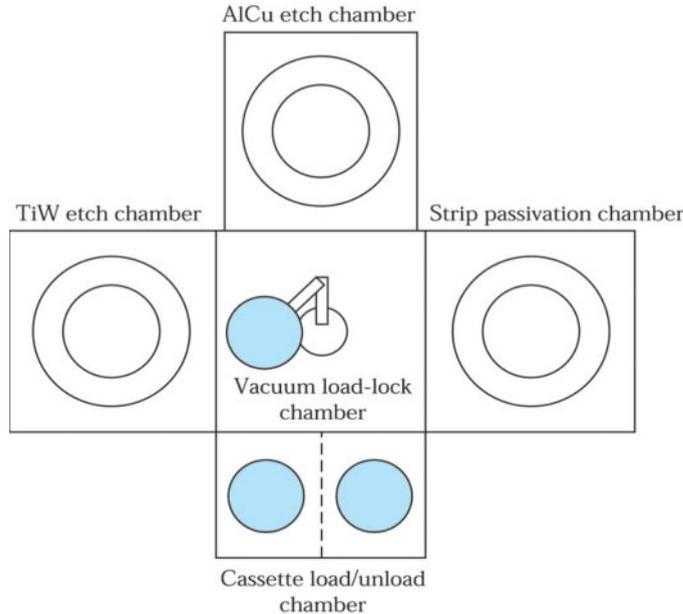


Fig. 28 Cluster reactive ion etch tool for multilayer metal (TiW/AlCu/TiW) interconnect etching.²

load-locked chamber so that the process chamber need not be vented under normal use. This enhanced uniformity, combined with automatic endpoint detection and microprocessor control, can also provide good process control.

A drawback of single-wafer etchers is that they must etch at higher rates to compete with the throughput of batch etchers. This constraint forces commercial single-wafer etchers to operate at higher rf power densities and sometimes higher pressures, where process control and selectivity are more difficult to achieve. For this reason, some manufactures offer hybrid reactors that combine a few single-wafer etchers in one machine.

Clustered Plasma Processing

Semiconductor wafers are processed in clean rooms to minimize exposure to ambient particulate contamination. As device dimensions shrink, particulate contamination becomes a more serious problem. To minimize particulate contamination, clustered plasma tools use a wafer handler to pass wafers from one process chamber to another in a vacuum environment. The clustered plasma processing tools can also increase throughput. Figure 28 shows the multilayer metal interconnect (TiW/AlCu/TiW) etching process with clustered tools of an AlCu etch chamber, a TiW etch chamber, and a strip passivation chamber. The clustered tools provide an economic advantage through their high chip yield because the wafer is exposed to less ambient contamination and is handled less.

13.4.5 Plasma Diagnostics and End-Point Control

Plasma Diagnostics

Most processing plasmas emit radiation from infrared to ultraviolet. A simple analytical technique is to measure the intensity of these emissions versus wavelength with the aid of optical emission spectroscopy (OES). Using observed spectral peaks, it is usually possible to determine the presence of neutral and ionic species by correlating these emissions with previously determined spectral series. Relative concentrations of the species can be obtained by correlating changes in intensity with the plasma parameter. The emission signal derived from the primary etchant or byproduct begins to rise or fall at the end of the etch cycle.

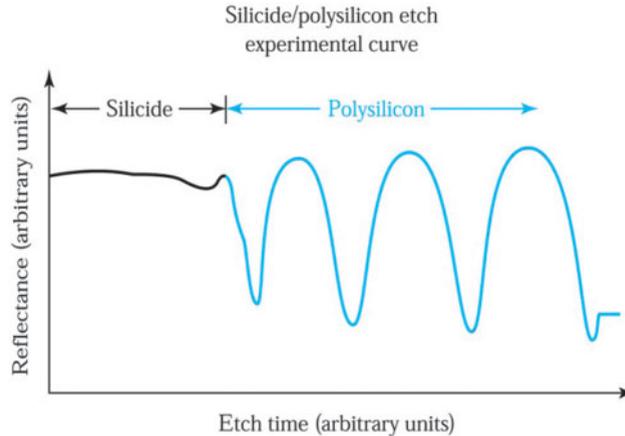


Fig. 29 The relative reflectance of the etching surface of a composite silicide/poly-Si layer. The end point of the etch is indicated by cessation of the reflective oscillation.

End-Point Control

Dry etching differs from wet chemical etching in that dry etching does not have enough etch selectivity to the underlying layer. Therefore, the plasma reactor must be equipped with a monitor that indicates when the etching process is to be terminated (i.e., an end point detection system). Laser interferometry of the wafer surface is used to continuously monitor etch rates and to determine the end point. During etching, the intensity of laser light reflected from a thin film surface oscillates because of the phase interference between the light reflected from the outer and inner interfaces of the etching layer. This layer must therefore be optically transparent or semitransparent to observe the oscillation. Figure 29 shows a typical signal from a silicide/polycrystalline Si gate etch. The period of the oscillation is related to the change in film thickness by

$$d / 2\bar{n} \quad (17)$$

where Δd is the change in film thickness for one period of reflected light, λ is the wavelength of the laser light, and \bar{n} is the refractive index of the etching layer. For example, Δd for polysilicon is 80 nm, measured by using a helium-neon laser source for which $\lambda = 632.8$ nm. The end point of the etch is indicated by the cessation of the reflection oscillation.

13.4.6 Etching Chemistries and Applications

Besides the etching equipment, etch chemistry also plays a critical role in the performance of etch processes. Table 2 lists some etch chemistries for different etch processes.

Silicon Trench Etching

As device feature size decreases, a corresponding decrease is needed in the wafer surface area occupied by the isolation between circuit elements and the storage capacitor of a DRAM cell. This surface area can be reduced by etching trenches into the silicon substrate and filling them with suitable dielectric or conductive materials. Deep trenches, usually with depths greater than 5 μm , are used mainly for forming storage capacitors. Shallow trenches, usually with depths less than 1 μm , are often used for isolation.

Chlorine-based and bromine-based chemistries have a high silicon etch rate and high etch selectivity to the silicon dioxide mask. The combination $\text{HBr} + \text{NF}_3 + \text{SF}_6 + \text{O}_2$ gas mixtures is used to form a trench capacitor with a depth of $\sim 7 \mu\text{m}$. It is also used for shallow trench isolation etching. Aspect-ratio-dependent

TABLE 2 ETCH CHEMISTRIES OF DIFFERENT ETCH PROCESSES

| Material being etched | Etching chemistry |
|--|--|
| Deep Si trench | HBr/NF ₃ /O ₂ /SF ₆ |
| Shallow Si trench | HBr/Cl ₂ /O ₂ |
| Poly Si | HBr/Cl ₂ /O ₂ , HBr/O ₂ , BCl ₃ /Cl ₂ , SF ₆ |
| Al | BCl ₃ /Cl ₂ , SiCl ₄ /Cl ₂ , HBr/Cl ₂ |
| AlSiCu | BCl ₃ /Cl ₂ /N ₂ |
| W | SF ₆ only NF ₃ /Cl ₂ |
| TiW | SF ₆ only |
| WSi ₂ , TiSi ₂ , CoSi ₂ | CCl ₂ F ₂ /NF ₃ , CF ₄ /Cl ₂ , Cl ₂ /N ₂ /C ₂ F ₆ |
| SiO ₂ | CF ₄ /CHF ₃ /Ar, C ₂ F ₆ , C ₃ F ₈ , C ₄ F ₈ /CO, C ₅ F ₈ , CH ₂ F ₂ |
| Si ₃ N ₄ | CHF ₃ /O ₂ , CH ₂ F ₂ , CH ₂ CHF ₂ , SF ₆ /He |

etching (i.e., variation in etch rate with aspect ratio) is often observed in submicron-deep silicon trench etching, caused by limited ion and neutral transport within the trench. Trenches with large aspect ratios are etched more slowly than trenches with small aspect ratios.

Polysilicon and Polycide Gate Etching

Polysilicon or polycide (i.e., low-resistance metal silicides over polysilicon) is usually used as a gate material for MOS devices. Anisotropic etching and high etch selectivity to the gate oxide are the most important requirements for gate etching. For example, the selectivity required in 1G DRAM is more than 150 (i.e., the ratio of etch rates for polycide and gate oxide is 150:1). Achieving high selectivity and etch anisotropy at the same time is difficult for most ion-enhanced etching processes. Therefore, multistep processing is used in which different etch steps in the process are optimized for etch anisotropy and selectivity. On the other hand, the trend in plasma technology for anisotropic etching and high selectivity is to utilize a low-pressure, high-density plasma using relatively low power. Most chlorine-based and bromine-based chemistries can be used for gate etching to achieve the required etch anisotropy and selectivity.

Dielectric Etching

The patterning of dielectrics, especially silicon dioxide and silicon nitride, is a key process in the manufacture of modern semiconductor devices. Because of their higher bonding energies, dielectric etching requires aggressive ion-enhanced, fluorine-based plasma chemistry. Vertical profiles are achieved by sidewall passivation as discussed in Section 13.4.2, typically by introducing a carbon-containing fluorine species to the plasma (e.g., CF₄, CHF₃, C₄F₈). High ion-bombardment energies are required to remove this polymer layer from the oxide, as well as to mix the reactive species into the oxide surface to form SiF_x products.

A low-pressure, high-density plasma is advantageous for aspect-ratio-dependent etching. However, the high-density plasma etchers (HDP, e.g., ICP and ECR) generate high-temperature electrons and subsequently produce a high degree of dissociation of ions and radicals, far more active radicals and ions than RIE or MERIE plasmas. In particular, a high F concentration worsens the selectivity to silicon. Various methods have been tried to enhance the selectivity in the high-density plasma. A parent gas with a high C/F ratio, such as C₂H₆, C₄H₈, or C₅H₈, has been successfully tried. Also, other methods to scavenge F radicals have been developed.²²

Interconnect Metal Etching

Etching of a metallization layer is a very important step in IC fabrication. Aluminum, copper, and tungsten are the most popular materials used for interconnection. These materials usually require anisotropic etching. Chlorine-based (e.g., Cl₂/BCl₃ mixture) chemistry has a very high chemical etch rate with aluminum and tends to produce an

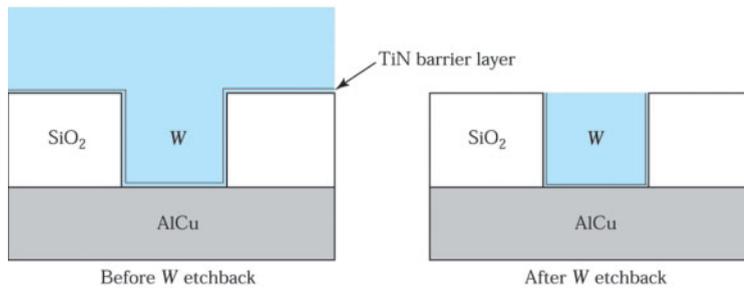


Fig. 30 Formation of tungsten plug in a contact hole by depositing blanket low-pressure chemical-vapor deposition W and then using reaction ion etching etchback.

undercut during etching. Carbon-containing gas (e.g., CHF_3) or N_2 is added to form sidewall passivation during aluminum etching to obtain anisotropic etching.

Copper has attracted much attention because of its low resistivity ($\sim 1.7 \mu\text{ohm-cm}$) and superior resistance to electromigration compared with Al or Al alloys. However, because of the low volatility of copper halides, plasma etching at room temperature is difficult. Process temperatures higher than 200°C are required to etch copper films. Therefore, the damascene process is used to form Cu interconnection without dry etching. Damascene processing, as discussed in Chapter 12, involves the creation of interconnect lines by first etching a trench or canal in a planar dielectric layer and then filling that trench with metal, such as aluminum or copper. In dual damascene processing shown in Fig. 26 in Section 12.5.4, Chapter 12, a second level is involved where a series of holes (i.e., contacts or vias) are etched and filled in addition to the trench. After filling, the metal and dielectric are planarized by chemical-mechanical processing (CMP). The advantage of damascene processing is that it eliminates the need for metal etch. This is an important concern as the industry moves from aluminum to copper interconnections.

Low-pressure CVD (LPCVD) tungsten (W) has been widely used for filling contact holes and first-level metallization because of its excellent deposition conformability. Both fluorine- and chlorine-based chemistries etch W and form volatile etch products. An important tungsten etch process is the blanket W etchback to form a W plug. The blanket LPCVD W is deposited on top of a TiN barrier layer, as shown in Fig. 30. A two-step process is usually used. First, 90% of the W is etched at a high etch rate, and then the etch rate is reduced to remove the remaining W with an etchant with a high W-to-TiN selectivity.

► SUMMARY

The continued growth of the semiconductor industry is a direct result of the ability to transfer smaller and smaller circuit patterns onto semiconductor wafers. The two major processes for transferring patterns are lithography and etching.

Currently, the vast majority of lithographic equipment is optical systems. The primary factor limiting resolution in optical lithography is diffraction. However, because of advancements in excimer lasers, photoresist chemistry, and resolution enhancement techniques such as the PSM, OPC and immersion technique, optical lithography will remain the mainstream technology, at least to the 32 nm generation.

Electron-beam lithography is the technology of choice for mask making and nanofabrication, in which new device concepts are explored. Other lithographic processing technologies are EUV and ion-beam lithography. Although all these have 100 nm or better resolution, each process has its own limitation: proximity effect in electron-beam lithography, mask blank production difficulties in EUV lithography, and stochastic space charge in ion-beam lithography.

At the present time, no obvious successor to optical lithography can be identified unambiguously. However, a mix-and-match approach can take advantage of the unique features of each lithographic process to improve resolution and to maximize throughput.

Wet chemical etching is used extensively in semiconductor processing. It is particularly suitable for blanket etching. We have discussed wet chemical etching processes for silicon and gallium arsenide, insulators, and metal

interconnections. The undercutting of the layer underneath the mask has resulted in loss of resolution in the etched pattern.

Dry etching methods are used to achieve high-fidelity pattern transfer. We have considered plasma fundamentals and various dry-etching systems, which have grown from relatively simple, parallel-plate configurations to complex chambers with multiple-frequency generators and a variety of process-control sensors.

The challenges for future etching technology are high etch selectivity, better critical-dimension control, high aspect-ratio-dependent etching, and low plasma-induced damage. Low-pressure, high-density plasma reactors are necessary to meet these requirements. As processing evolves from 300 mm to even larger wafers, continued improvements are required for etch uniformity within the wafer. New gas chemistries must be developed to provide the improved selectivity necessary for advanced integration circuits.

► REFERENCES

1. For a more detailed discussion on lithography, see (a) R. Doering and Y. Nishi, Ed., *Handbook of Semiconductor Manufacturing Technology*, 2nd Ed., CRC Press, Florida, 2008. (b) K. Nakamura, "Lithography," in C. Y. Chang and S. M. Sze, Eds., *ULSI Technology*, McGraw-Hill, New York, 1996. (c) P. Rai-Choudhurg, *Handbook of Microlithography, Micromachining, and Microfabrication*, Vol. 1, SPIE, Washington, DC, 1997. (d) D. A. McGillis, "Lithography," in S. M. Sze, Ed., *VLSI Technology*, McGraw-Hill, New York, 1983.
2. For a more detailed discussion of etching, see Y. J. T. Liu, "Etching," in C. Y. Chang and S. M. Sze, Eds., *ULSI Technology*, McGraw-Hill, New York, 1996.
3. J. M. Duffalo and J. R. Monkowski, "Particulate Contamination and Device Performance," *Solid State Technol.* **27**, 3, 109 (1984).
4. H. P. Tseng and R. Jansen, "Cleanroom Technology," in C. Y. Chang and S. M. Sze, Eds., *ULSI Technology*, McGraw-Hill, New York, 1996.
5. M. C. King, "Principles of Optical Lithography," in N. G. Einspruch, Ed., *VLSI Electronics*, Vol. 1, Academic, New York, 1981.
6. J. H. Bruning, "A Tutorial on Optical Lithography," in D. A. Doane, et al., Eds., *Semiconductor Technology*, Electrochemical Soc., Pennington, 1982.
7. R. K. Watts and J. H. Bruning, "A Review of Fine-Line Lithographic Techniques: Present and Future," *Solid State Technol.*, **24**, 5, 99 (1981).
8. W. C. Till and J. T. Luxon, *Integrated Circuits, Materials, Devices, and Fabrication*, Prentice-Hall, Englewood Cliffs, NJ, 1982.
9. M. D. Levenson, N. S. Viswanathan, and R. A. Simpson, "Improving Resolution in Photolithography with a Phase-Shift Mask," *IEEE Trans. Electron Devices*, **ED-29**, 18–28 (1982).
10. D. P. Kern et al., "Practical Aspects of Microfabrication in the 100-nm Region," *Solid State Technol.*, **27**, 2, 127 (1984).
11. J. A. Reynolds, "An Overview of e-Beam Mask-Making," *Solid State Technol.*, **22**, No. 8, 87 (1979).
12. Y. Sameda, et al. "Electron-Beam Cell Projection Lithography: Its Accuracy and Its Throughput," *J. Vac. Sci. Technol.*, **B12** (6), 3399 (1994).
13. W. L. Brown, T. Venkatesan, and A. Wagner, "Ion Beam Lithography," *Solid State Technol.*, **24**, 8, 60 (1981).

14. D. S. Kyser and N. W. Viswanathan, "Monte Carlo Simulation of Spatially Distributed Beams in Election-Beam Lithography," *J. Vac. Sci. Technol.*, **12**, 1305 (1975).
15. Charles Gwyn et al., *Extreme Ultraviolet Lithography-White Paper*, Sematech, Next-Generation Lithography Workshop, Colorado Springs, Dec. 7–10, 1998.
16. L. Karapiperis et al., "Ion Beam Exposure Profiles in PMMA-Computer Simulation," *J. Vac. Sci. Technol.*, **19**, 1259 (1981).
17. H. Robbins and B. Schwartz, "Chemical Etching of Silicon II, the System HF, HNO₃, H₂O and HC₂H₃O₂," *J. Electrochem. Soc.*, **107**, 108 (1960).
18. K. E. Bean, "Anisotropic Etching in Silicon," *IEEE Trans. Electron Devices*, **ED-25**, 1185 (1978).
19. S. Iida and K. Ito, "Selective Etching of Gallium Arsenide Crystal in H₂SO₄-H₂O₂-H₂O System," *J. Electrochem. Soc.*, **118**, 768 (1971).
20. E. C. Douglas, "Advanced Process Technology for VLSI Circuits," *Solid State Technol.*, **24**, 5, 65 (1981).
21. M. Armacost et al., "Plasma-Etching Processes for ULSI Semiconductor Circuits," *IBM J. Res. Dev.*, **43**, 39 (1999).
22. C. O. Jung et al., "Advanced Plasma Technology in Microelectronics," *Thin Solid Films*, **341**, 112, (1999).

► **PROBLEMS (* DENOTES DIFFICULT PROBLEMS)**

FOR SECTION 13.1 OPTICAL LITHOGRAPHY

1. For a class-100 clean room, find the number of dust particles per cubic meter with particle sizes (a) between 0.5 and 1 μm , (b) between 1 and 2 μm , and (c) above 2 μm .
2. Find the final yield for a nine-mask-level process in which the average fatal defect density per cm^2 is 0.1 for four levels, 0.25 for four levels, and 1.0 for one level. The chip area is 50 mm^2 .
3. An optical lithographic system has an exposure power of 0.3 mW/cm^2 . The required exposure energy for a positive photoresist is 140 mJ/cm^2 and for a negative photoresist is 9 mJ/cm^2 . Assuming negligible times for loading and unloading wafers, compare the wafer throughput for positive photoresist and negative photoresist.
4. (a) For an ArF excimer laser 193 nm optical lithographic system with NA = 0.65, $k_1 = 0.60$, and $k_2 = 0.50$, what are the theoretical resolution and depth of focus for this equipment? (b) What can we do in practice to adjust NA, k_1 , and k_2 parameters to improve resolution? (c) What parameter does the phase-shift mask (PSM) technique change to improve resolution?
5. The plots in Fig. 9 are called *response curves* in microlithography. (a) What are the advantages and disadvantages of using resists with high γ values? (b) Conventional resists cannot be used for 248 nm or 193 nm lithography. Why not?

FOR SECTION 13.2 NEXT-GENERATION LITHOGRAPHIC METHODS

6. (a) Explain why a shaped beam promises higher throughput than a Gaussian beam in e-beam lithography. (b) How can alignment be performed for e-beam lithography?
7. Why has the operating mode of optical lithographic systems evolved from proximity printing to 1:1 projection printing and finally to 5:1 projection step-and-repeat?

FOR SECTION 13.3 WET CHEMICAL ETCHING

8. If the mask and the substrate cannot be etched by a particular etchant, sketch the edge profile of an isotropically etched feature in a film of thickness h_f for (a) etching just to completion, (b) 100% overetch, and (c) 200% overetch.
9. A $\langle 100 \rangle$ -oriented silicon crystal is etched in a KOH solution through a $1.5 \mu\text{m} \times 1.5 \mu\text{m}$ window defined in silicon dioxide. The etch rate normal to (100)-planes is $0.6 \mu\text{m}/\text{min}$. The etch rate ratios are 100:16:1 for the (100):(110):(111)-planes. Show the etched profile after 20 seconds, 40 seconds, and 60 seconds.
10. Repeat the previous problem. a $\langle \bar{1}10 \rangle$ -oriented silicon is etched with a thin SiO_2 mask in KOH solution. Show the etched pattern profiles on $\langle \bar{1}10 \rangle$ -Si.
11. A $\langle 100 \rangle$ -oriented silicon wafer 150 mm in diameter is $625 \mu\text{m}$ thick. The wafer has $1000 \mu\text{m} \times 1000 \mu\text{m}$ ICs on it. The IC chips are to be separated by orientation-dependent etching. Describe two methods for doing this and calculate the fraction of the surface area that is lost in these processes.

FOR SECTION 13.4 DRY ETCHING

- *12. The average distance traveled by particles between collisions is called the mean free path (λ), $\lambda \cong 5 \times 10^{-3}/P(\text{cm})$, where P is pressure in Torr. In typical plasmas of interest, the chamber pressure ranges from 1 Pa to 150 Pa. What are the corresponding density of gas molecules (cm^{-3}) and the mean free path?
13. Fluorine (F) atoms etch Si at a rate given by Etch Rate (nm/min) = $2.86 \times 10^{-13} n_F \times T^{1/2} e^{-E_a/RT}$ where n_F is the concentration of F atoms (cm^{-3}), T the temperature (K), and E_a and R the activation energy (2.48 kcal/mol) and gas constant (1.987 cal-K), respectively. If n_F is 3×10^{15} , calculate the etch rate of Si at room temperature.
14. Repeat the previous problem. SiO_2 etched by F atoms can also be expressed by Etch rate (nm/min) = $0.614 \times 10^{-13} n_F \times T^{1/2} e^{-E_a/RT}$ where n_F is 3×10^{15} (cm^{-3}) and E_a is 3.76 kcal/mol. Calculate the etch rate of SiO_2 and etch selectivity of SiO_2 over Si at room temperature.
15. A multiple-step etch process is required for etching a polysilicon gate with thin gate oxide. How do you design an etch process that has no micromasking, has an anisotropic etch profile, and is selective to thin gate oxide?
16. Find the etch selectivity required to etch a 400-nm polysilicon layer without removing more than 1 nm of its underlying gate oxide, assuming that the polysilicon is etched with a process having a 10% etch-rate uniformity.
17. A $1 \mu\text{m}$ Al film is deposited over a flat field oxide region and patterned with photoresist. The metal is then etched with a mixture of BCl_3/Cl_2 gases at a temperature of 70°C in a Helicon etcher. The selectivity of Al over photoresist is maintained at 3. Assuming a 30% overetch, what is the minimum photoresist thickness required to ensure that the top metal surface is not attacked?
18. In an ECR plasma, a static magnetic field B forces electrons to circulate around the magnetic field lines at an angular frequency, ω_e , that is given by

$$\omega_e = qB/m_e,$$
 where q is the electronic charge and m_e the electron mass. If the microwave frequency is 2.45 GHz, what is the required magnetic field?
19. What are the major distinctions between the traditional reactive ion etching and high-density plasma etching (ECR, ICP, etc.)?
20. Describe how to eliminate corrosion issues in Al lines after etching with chlorine-based plasma.

Impurity Doping

- ▶ 14.1 BASIC DIFFUSION PROCESS
 - ▶ 14.2 EXTRINSIC DIFFUSION
 - ▶ 14.3 DIFFUSION-RELATED PROCESSES
 - ▶ 14.4 RANGE OF IMPLANTED IONS
 - ▶ 14.5 IMPLANT DAMAGE AND ANNEALING
 - ▶ 14.6 IMPLANTATION-RELATED PROCESSES
 - ▶ SUMMARY
-

Impurity doping is the introduction of controlled amounts of impurity dopants into semiconductor materials. The practical use of impurity doping is primarily to change the electrical properties of the semiconductors. *Diffusion* and *ion implantation* are the two key methods of impurity doping. Until the early 1970s, impurity doping was done mainly by diffusion at elevated temperatures, as shown in Fig. 1a. In this method the dopant atoms are placed on or near the surface of the wafer by deposition from the gas phase of the dopant or by using doped-oxide sources. The doping concentration decreases monotonically from the surface, and the profile of the dopant distribution is determined mainly by the temperature and diffusion time.

Since the early 1970s, many doping operations have been performed by ion implantation, as shown in Fig. 1b. In this process the dopant ions are implanted into the semiconductor by means of an ion beam. The doping concentration has a peak distribution inside the semiconductor and the profile of the dopant distribution is determined mainly by the ion mass and the implanted-ion energy. Both diffusion and ion implantation are used in fabricating discrete devices and integrated circuits because these processes generally complement each other.^{1,2} For example, diffusion is used to form a deep junction (e.g., a twin well in CMOS), whereas ion implantation is used to form a shallow junction (e.g., a source/drain junction of a MOSFET).

Specifically, we cover the following topics:

- The movement of impurity atoms in the crystal lattice under high temperature and high concentration-gradient conditions.
- Impurity profiles for constant diffusivity and concentration-dependent diffusivity.
- The impact of lateral diffusion and impurity redistribution on device characteristics.
- The process and advantages of ion implantation.
- Ion distributions in the crystal lattice and how to remove lattice damage caused by ion implantation.
- Implantation-related processes such as masking, high-energy implantation, and high-current implantation.

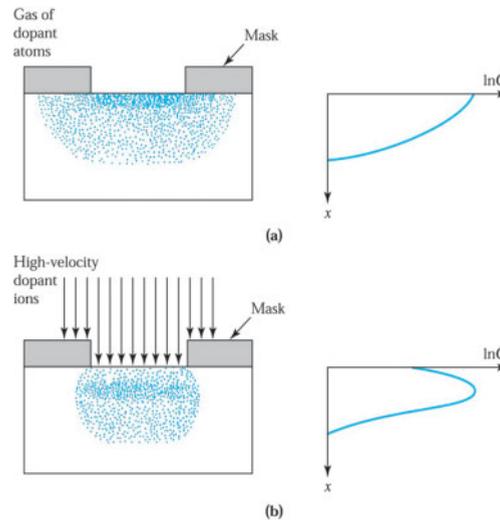


Fig. 1 Comparison of (a) diffusion and (b) ion-implantation techniques for selective introduction of dopants into the semiconductor substrate.

► 14.1 BASIC DIFFUSION PROCESS

Diffusion of impurities is typically done by placing semiconductor wafers in a carefully controlled high-temperature quartz-tube furnace and passing a gas mixture that contains the desired dopant through it. The temperature usually ranges between 800° and 1200°C for silicon and 600° and 1000°C for gallium arsenide. The number of dopant atoms that diffuse into the semiconductor is related to the partial pressure of the dopant impurity in the gas mixture.

For diffusion in silicon, boron is the most popular dopant for introducing a *p*-type impurity, whereas arsenic and phosphorus are used extensively as *n*-type dopants. These three elements are highly soluble in silicon, as they have solubilities above $5 \times 10^{20} \text{ cm}^{-3}$ in the diffusion temperature range. These dopants can be introduced in several ways, including solid sources (e.g., BN for boron, As_2O_3 for arsenic, and P_2O_5 for phosphorus), liquid sources (BBr_3 , AsCl_3 , and POCl_3), and gaseous sources (B_2H_6 , AsH_3 , and PH_3). However, liquid sources are most commonly used. A schematic diagram of the furnace and gas flow arrangement for a liquid source is shown in Fig. 2. This arrangement is similar to that used for thermal oxidation. An example of the chemical reaction for phosphorus diffusion using a liquid source is



The P_2O_5 forms a glass on a silicon wafer and is then reduced to phosphorus by silicon,



the phosphorus is released and diffuses into the silicon and Cl_2 is vented.

For diffusion in gallium arsenide, the high vapor pressure of arsenic requires special methods to prevent the loss of arsenic by decomposition or evaporation.² These methods include diffusion in sealed ampules with an overpressure of arsenic and diffusion in an open-tube furnace with a doped-oxide capping layer (e.g., silicon nitride). Most of the studies on *p*-type diffusion have been confined to the use of zinc in the forms of Zn-Ga-As alloys and ZnAs_2 for the sealed-ampule approach or ZnO-SiO_2 for the open-tube approach. The *n*-type dopants in gallium arsenide include selenium and tellurium.

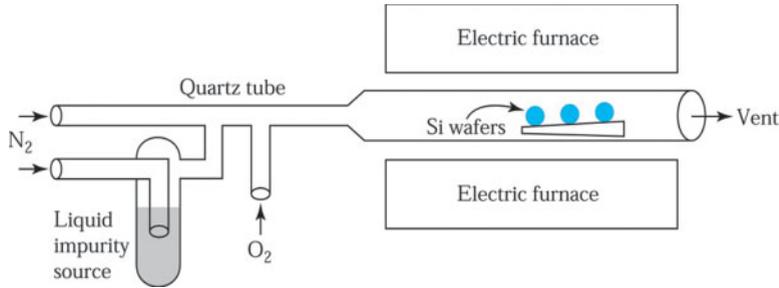


Fig. 2 Schematic diagram of a typical open-tube diffusion system.

14.1.1 Diffusion Equation

Diffusion in a semiconductor can be visualized as atomic movement of the diffusant (dopant atoms) in the crystal lattice by vacancies or interstitials. Fig.3 shows the two basic atomic-diffusion models in a solid.^{1,3} The open circles represent the host atoms occupying the equilibrium lattice positions and the solid dots represent impurity atoms. At elevated temperatures the lattice atoms vibrate around the equilibrium lattice sites. There is a finite probability that a host atom acquires sufficient energy to leave the lattice site and become an interstitial atom, thereby creating a vacancy. When a neighboring impurity atom migrates to the vacancy site as illustrated in Fig. 3a, the mechanism is called *vacancy diffusion*. If an interstitial atom moves from one place to another without occupying a lattice site (Fig. 3b), the mechanism is *interstitial diffusion*. An atom smaller than the host atom often moves interstitially.

In addition, there is extended interstitial diffusion, sometimes called *interstitialcy diffusion*. The interstitial host atom (self-interstitial) pushes the substitutional impurity atom into an interstitial site. Subsequently, the impurity atom displaces another host atom and creates a new self-interstitial. Then the process is repeated. Interstitialcy diffusion is faster than substitutional diffusion. Vacancy and interstitialcy diffusion are considered the dominant mechanisms for diffusion of P, B, As, and Sb in silicon. Phosphorus and boron diffuse via a dual (vacancy and interstitialcy) mechanism, with the interstitialcy component dominating. Arsenic and antimony diffuse predominately via a vacancy mechanism.¹

The basic diffusion process of impurity atoms is similar to that of charge carriers (electrons and holes), discussed in Chapter 2. Accordingly, we define a flux F as the number of dopant atoms passing through a unit area in a unit time and C as the dopant concentration per unit volume. From Eq. 27 in Chapter 2, we have

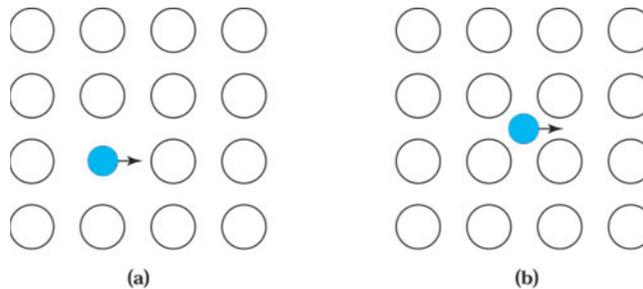


Fig. 3 Atomic diffusion mechanisms for a two-dimensional lattice.^{1,3} (a) Vacancy mechanism; (b) interstitial mechanism.

$$F = -D \frac{\partial C}{\partial x}, \quad (3)$$

where we have substituted C for the carrier concentration and the proportionality constant D is the diffusion coefficient or diffusivity. Note that the basic driving force of the diffusion process is the concentration gradient dC/dx . The flux is proportional to the concentration gradient, and the dopant atoms will move (diffuse) away from a high-concentration region toward a lower-concentration region.

If we substitute Eq. 3 into the one-dimensional continuity equation, Eq. 56 in Chapter 2, under the condition that no materials are formed or consumed in the host semiconductor (i.e., $G_n = R_n = 0$), we obtain

$$\frac{\partial C}{\partial t} = -\frac{\partial F}{\partial x} = \frac{\partial}{\partial x} \left(D \frac{\partial C}{\partial x} \right). \quad (4)$$

When the concentration of the dopant atoms is low, the diffusion coefficient can be considered independent of doping concentration, and Eq. 4 becomes

$$\boxed{\frac{\partial C}{\partial t} = D \frac{\partial^2 C}{\partial x^2}}. \quad (5)$$

Equation 5 is often referred to as *Fick's diffusion equation*.

Figure 4 shows the measured diffusion coefficients for low concentrations of various dopant impurities in silicon and gallium arsenide.^{4,5} The logarithm of the diffusion coefficient plotted against the reciprocal of the absolute temperature gives a straight line in most cases. This implies that over the temperature range, the diffusion coefficients can be expressed as

$$\boxed{D = D_0 \exp\left(\frac{-E_a}{kT}\right)}, \quad (6)$$

where D_0 is the diffusion coefficient in cm^2/s extrapolated to infinite temperature and E_a is the activation energy in eV.

For the interstitial diffusion model, E_a is related to the energies required to move dopant atoms from one interstitial site to another. The values of E_a are found to be between 0.5 and 2 eV in both silicon and gallium arsenide. For the vacancy diffusion model, E_a is related to both the energies of motion and the energies of formation of vacancies. Thus, E_a for vacancy diffusion is larger than that for interstitial diffusion; usually between 3 and 5 eV.

For fast diffusants, such as Cu in Si and GaAs, shown in the upper portion of Fig. 4a and 4b, the measured activation energies are less than 2 eV, and interstitial atomic movement is the dominant diffusion mechanism. For slow diffusants, such as As in Si and GaAs, shown in the lower portion of Fig. 4a and 4b, E_a is larger than 3 eV, and vacancy diffusion is the dominant mechanism. For interstitialcy-dominated diffusion, such as P in silicon, E_a is also larger than 3 eV, but the diffusion coefficient is four times greater than As over the temperature range shown in Fig. 4a.

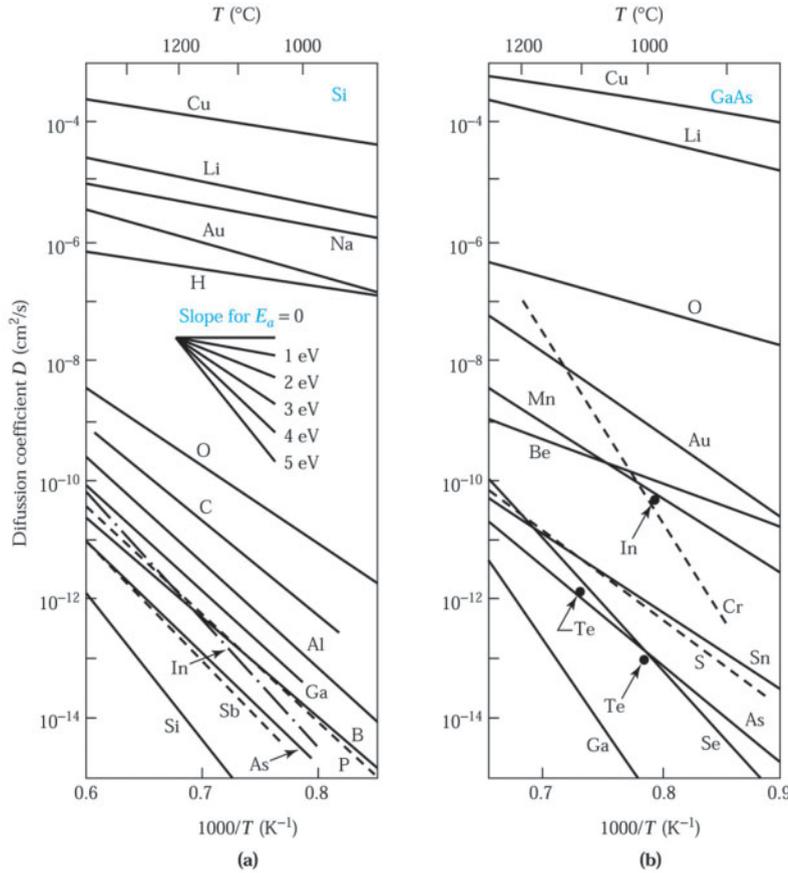


Fig. 4 Diffusion coefficient (also called diffusivity) as a function of the reciprocal of temperature for (a) silicon and (b) gallium arsenide.^{4,5}

14.1.2 Diffusion Profiles

The diffusion profile of the dopant atoms is dependent on the initial and boundary conditions. In this subsection we consider two important cases, namely, constant-surface-concentration diffusion and constant-total-dopant diffusion. In the first case, impurity atoms are transported from a vapor source onto the semiconductor surface and diffuse into the semiconductor wafers. The vapor source maintains a constant level of surface concentration during the entire diffusion period. In the second case, a fixed amount of dopant is deposited onto the semiconductor surface and is subsequently diffused into the wafers.

Constant-Surface-Concentration Diffusion

The initial condition at $t = 0$ is

$$C(x, 0) = 0, \tag{7}$$

which states that the dopant concentration in the host semiconductor is initially zero. The boundary conditions are

$$C(0, t) = C_s \tag{8a}$$

and

$$C(\infty, t) = 0 \quad (8b)$$

where C_s is the surface concentration (at $x = 0$), which is independent of time. The second boundary condition states that at large distances from the surface there are no impurity atoms.

The solution of the diffusion equation (Eq. 5) that satisfies the initial and boundary conditions is given by⁶

$$C(x, t) = C_s \operatorname{erfc}\left(\frac{x}{2\sqrt{Dt}}\right), \quad (9)$$

where erfc is the complementary error function and \sqrt{Dt} is the diffusion length. The definition of erfc and some properties of the function are summarized in Table 1. The diffusion profile for the constant-surface-concentration condition is shown in Fig. 5a, where we plot, on both linear (upper) and logarithmic (lower) scales, the normalized concentration as a function of depth for three values of the diffusion length \sqrt{Dt} corresponding to three consecutive diffusion times and a fixed D for a given diffusion temperature. Note that as the time progresses, the dopant penetrates deeper into the semiconductor.

The total number of dopant atoms per unit area of the semiconductor is given by

$$Q(t) = \int_0^{\infty} C(x, t) dx. \quad (10)$$

TABLE 1 ERROR FUNCTION ALGEBRA

$$\operatorname{erf}(x) \equiv \frac{2}{\sqrt{\pi}} \int_0^x e^{-y^2} dy$$

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x)$$

$$\operatorname{erf}(0) = 0$$

$$\operatorname{erf}(\infty) = 1$$

$$\operatorname{erf}(x) \cong \frac{2}{\sqrt{\pi}} x \quad \text{for } x \ll 1$$

$$\operatorname{erfc}(x) \cong \frac{1}{\sqrt{\pi}} \frac{e^{-x^2}}{x} \quad \text{for } x \gg 1$$

$$\frac{d}{dx} \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} e^{-x^2}$$

$$\frac{d^2}{dx^2} \operatorname{erf}(x) = -\frac{4}{\sqrt{\pi}} x e^{-x^2}$$

$$\int_0^x \operatorname{erfc}(y') dy' = x \operatorname{erfc}(x) + \frac{1}{\sqrt{\pi}} (1 - e^{-x^2})$$

$$\int_0^{\infty} \operatorname{erfc}(x) dx = \frac{1}{\sqrt{\pi}}$$

Substituting Eq. 9 into Eq. 10 yields

$$Q(t) = \frac{2}{\sqrt{\pi}} C_s \sqrt{Dt} \cong 1.13 C_s \sqrt{Dt}. \tag{11}$$

This expression can be interpreted as follows. The quantity $Q(t)$ represents the area under one of the diffusion profiles of the linear plot in Fig. 5a. These profiles can be approximated by triangles with height C_s and base $2\sqrt{Dt}$. This leads to $Q(t) \cong C_s \sqrt{Dt}$, which is close to the exact result obtained from Eq. 11.

A related quantity is the gradient of the diffusion profile dC/dx . The gradient can be obtained by differentiating Eq. 9:

$$\left. \frac{dC}{dx} \right|_{x,t} = -\frac{C_s}{\sqrt{\pi Dt}} e^{-x^2/4Dt}. \tag{12}$$

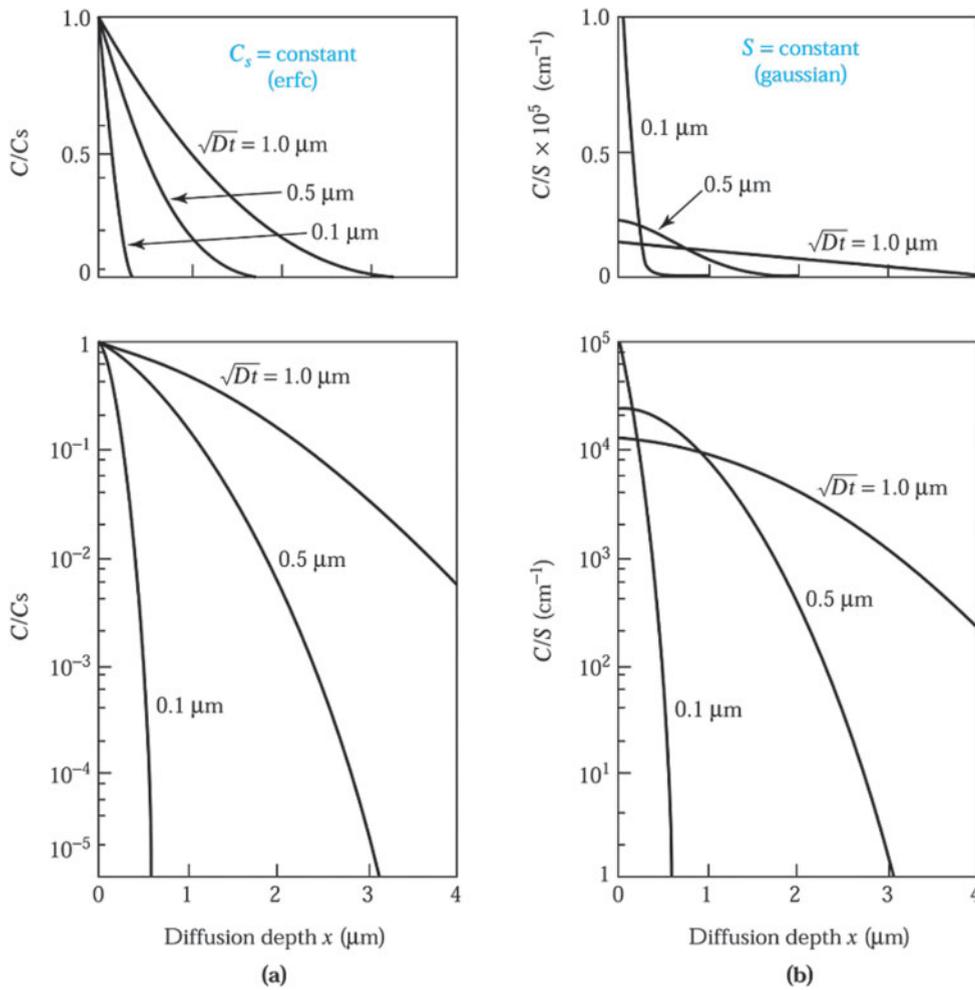


Fig. 5 Diffusion profiles. (a) Normalized complementary error function versus distance for successive diffusion times. (b) Normalized Gaussian function versus distance.

► **EXAMPLE 1**

For a boron diffusion in silicon at 1000°C, the surface concentration is maintained at 10^{19} cm^{-3} and the diffusion time is 1 hour. Find $Q(t)$ and the gradient at $x = 0$ and at a location where the dopant concentration reaches 10^{15} cm^{-3} .

SOLUTION The diffusion coefficient of boron at 1000°C, as obtained from Fig. 4, is about $2 \times 10^{-14} \text{ cm}^2/\text{s}$, so that the diffusion length is

$$\begin{aligned}\sqrt{Dt} &= \sqrt{2 \times 10^{-14} \times 3600} = 8.48 \times 10^{-6} \text{ cm}, \\ Q(t) &= 1.13 C_s \sqrt{Dt} = 1.13 \times 10^{19} \times 8.48 \times 10^{-6} = 9.5 \times 10^{13} \text{ atoms/cm}^2, \\ \left. \frac{dC}{dx} \right|_{x=0} &= -\frac{C_s}{\sqrt{\pi Dt}} = \frac{-10^{19}}{\sqrt{\pi} \times 8.48 \times 10^{-6}} = -6.7 \times 10^{23} \text{ cm}^{-4}.\end{aligned}$$

When $C = 10^{15} \text{ cm}^{-3}$, the corresponding distance x_j is given by Eq. 9, or

$$\begin{aligned}x_j &= 2\sqrt{Dt} \operatorname{erfc}^{-1}\left(\frac{10^{15}}{10^{19}}\right) = 2\sqrt{Dt} (2.75) = 4.66 \times 10^{-5} \text{ cm} = 0.466 \text{ } \mu\text{m}, \\ \left. \frac{dC}{dx} \right|_{x=0.466 \text{ } \mu\text{m}} &= -\frac{C_s}{\sqrt{\pi Dt}} e^{-x_j^2/4Dt} = -3.5 \times 10^{20} \text{ cm}^{-4}.\end{aligned}$$

Constant–Total-Dopant Diffusion

For this case, a fixed (or constant) amount of dopant is deposited onto the semiconductor surface in a thin layer and the dopant subsequently diffuses into the semiconductor. The initial condition is the same as in Eq. 7. The boundary conditions are

$$\int_0^{\infty} C(x, t) dx = S \quad (13a)$$

and

$$C(\infty, t) = 0 \quad (13b)$$

where S is the total amount of dopant per unit area.

The solution of the diffusion equation, Eq. 5, that satisfies the above conditions is

$$\boxed{C(x, t) = \frac{S}{\sqrt{\pi Dt}} \exp\left(-\frac{x^2}{4Dt}\right)} \quad (14)$$

This expression is the Gaussian distribution. Since the dopant will move into the semiconductor as time increases, to keep the total dopant S constant, the surface concentration must decrease. This is indeed the case, since the surface concentration is given by Eq. 14 with $x = 0$:

$$C_s(t) = \frac{S}{\sqrt{\pi Dt}}. \quad (15)$$

Figure 5b shows the dopant profile for a Gaussian distribution where we plot the normalized concentration (C/S) as a function of the distance for three increasing diffusion lengths. Note the reduction of the surface

concentration as the diffusion time increases. The gradient of the diffusion profile is obtained by differentiating Eq. 14 and is

$$\left. \frac{dC}{dx} \right|_{x,t} = \frac{xS}{2\sqrt{\pi}(Dt)^{3/2}} e^{-x^2/4Dt} - \frac{x}{2Dt} C(x,t). \quad (16)$$

The gradient (or slope) is zero at $x = 0$ and at $x = \infty$ and the maximum gradient occurs at $x = \sqrt{2Dt}$.

In integrated-circuit processing, a two-step diffusion process is commonly used in which a *predeposition* diffused layer is first formed under a constant-surface-concentration condition. This step is followed by a *drive-in* diffusion (also called *redistribution* diffusion) under a constant-total-dopant condition. For most practical cases, the diffusion length \sqrt{Dt} for the predeposition diffusion is much smaller than the diffusion length for the drive-in diffusion. Therefore, we can consider the predeposition profile as a delta function at the surface, and we can regard the extent of the penetration of the predeposition profile as negligibly small compared with that of the final profile that results from the drive-in step.

▶ EXAMPLE 2

Arsenic was predeposited by arsine gas and the resulting total amount of dopant per unit area is 1×10^{14} atoms/cm². How long would it take to drive the arsenic into a junction depth of 1 μm ? Assume a background doping of 1×10^{15} atoms/cm³ and a drive-in temperature of 1200°C. For As diffusion, $D_0 = 24$ cm²/s and $E_a = 4.08$ eV.

SOLUTION

$$D = D_0 \exp\left(\frac{-E_a}{kT}\right) = 24 \exp\left(\frac{-4.08}{8.614 \times 10^{-5} \times 1473}\right) = 2.602 \times 10^{-13} \text{ cm}^2/\text{s},$$

$$x_j^2 = 10^{-8} = 4Dt \ln\left(\frac{S}{C_B \sqrt{\pi Dt}}\right) = 1.04 \times 10^{-12} t \ln\left(\frac{1.106 \times 10^5}{\sqrt{t}}\right),$$

$$t \cdot \log t - 10.09t + 8350 = 0.$$

The solution to the above equation can be determined by the cross point of equation $y = t \cdot \log t$ and $y = 10.09t - 8350$. Therefore, $t = 1190$ seconds \cong 20 minutes. ◀

14.1.3 Evaluation of Diffused Layers

The results of a diffusion process can be evaluated by three measurements—the junction depth, the sheet resistance, and the dopant profile of the diffused layer. The junction depth can be delineated by cutting a groove into the semiconductor and etching the surface with a solution (e.g., 100 cm³ HF and a few drops of HNO₃ for silicon) that stains the *p*-type region darker than the *n*-type region, as illustrated in Fig. 6a. If R_0 is the radius of the tool used to form the groove, then the junction depth x_j is given by

$$x_j = \sqrt{R_0^2 - b^2} - \sqrt{R_0^2 - a^2}, \quad (17)$$

where a and b are indicated in the figure. In addition, if R_0 is much larger than a and b , then

$$x_j \cong \frac{a^2 - b^2}{2R_0}. \quad (18)$$

The junction depth x_j as illustrated in Fig. 6b is the position where the dopant concentration equals the substrate concentration C_B , or

$$C(x_j) = C_B. \quad (19)$$

Thus, if the junction depth and C_B are known, the surface concentration C_s and the impurity distribution can be calculated, provided the diffusion profile follows one or the other simple equation derived in Section 14.1.2.

The resistance of a diffused layer can be measured by the four-point probe technique described in Chapter 2. The *sheet resistance* R is related to the junction depth x_j , the carrier mobility μ (which is a function of the total impurity concentration), and the impurity distribution $C(x)$ by the following expression:⁷

$$R = \frac{1}{q \int_0^{x_j} \mu C(x) dx}. \quad (20)$$

For a given diffusion profile, the average resistivity $\bar{\rho} = Rx_j$ is uniquely related to the surface concentration C_s and the substrate-doping concentration for an assumed diffusion profile. Design curves relating C and $\bar{\rho}$ have been calculated for simple diffusion profiles, such as the erfc or Gaussian distribution.⁸ To use these curves correctly we must be sure that the diffusion profiles agree with the assumed profiles. For low concentrations and deep diffusions, the diffusion profiles generally can be represented by the aforementioned simple functions. However, as we discuss in the next section, for high concentrations and shallow diffusions, the diffusion profiles cannot be represented by these simple functions.

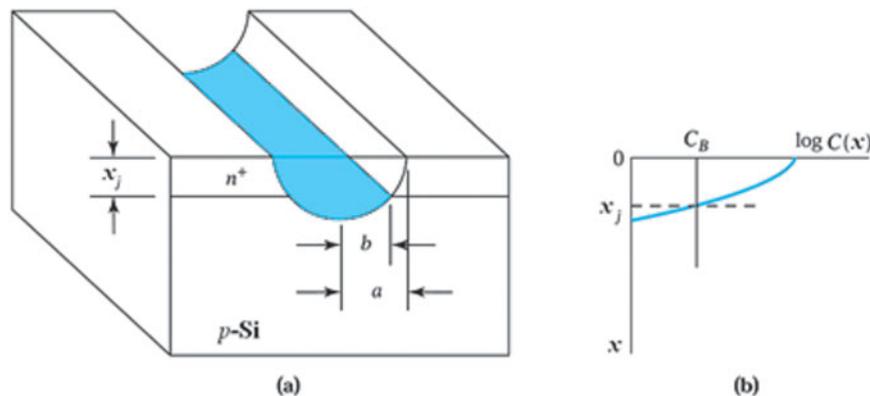


Fig. 6 Junction-depth measurement. (a) Grooving and staining; (b) position in which dopant and substrate concentration are equal.

The diffusion profile can be measured using the capacitance-voltage technique described in Chapter 7. The majority carrier profile, which is equal to the impurity profile if impurities are fully ionized, can be determined by measuring the reverse-bias capacitance of a p - n junction or a Schottky barrier diode as a function of the applied voltage. A more elaborate method is the secondary-ion-mass spectroscope (SIMS) technique, which measures the total impurity profile. In the SIMS technique, an ion beam sputters material off the surface of a semiconductor, and the ion component is detected and mass analyzed. This technique has high sensitivity to many elements, such as boron and arsenic, and is an ideal tool for providing the precision needed for profile measurements in high-concentration or shallow-junction diffusions.⁹

► 14.2 EXTRINSIC DIFFUSION

The diffusion profiles described in Section 14.1 are for constant diffusivities. These profiles occur when the doping concentration is lower than the intrinsic-carrier concentration $n_i(T)$ at the diffusion temperature. For example, at $T = 1000^\circ\text{C}$, n_i equals $5 \times 10^{18} \text{ cm}^{-3}$ for silicon and $5 \times 10^{17} \text{ cm}^{-3}$ for gallium arsenide. The diffusivity at low concentrations is often referred to as the intrinsic diffusivity $D_i(T)$. Doping profiles that have concentrations less than $n_i(T)$ are in the *intrinsic* diffusion region as indicated in the left side of Fig. 7. In this region, the resulting dopant profiles of sequential or simultaneous diffusions of n - and p -type impurities can be determined by superposition; that is, the diffusions can be treated independently. However, when the impurity concentration, including both the substrate and the dopant, is greater than $n_i(T)$, the semiconductor becomes extrinsic and the diffusivity is considered to be extrinsic. In the extrinsic diffusion region the diffusivity becomes concentration dependent.¹⁰ In the extrinsic diffusion region the diffusion profiles are more complicated, and there are interactions and cooperative effects among the sequential or simultaneous diffusions.

14.2.1 Concentration-Dependent Diffusivity

As mentioned previously, when a host atom acquires sufficient energy from the lattice vibration to leave its lattice site, a vacancy is created. The presence of a vacancy in a crystal results in four unsatisfied and distorted bonds. The electrons of these bonds may spill into the vacancy. A neutral vacancy will act as an acceptor by acquiring a negative charge, $V^0 + e^- = V^-$. Therefore, depending on the charges associated with a vacancy, we can have a neutral vacancy V^0 , an acceptor vacancy V^- , a double-charged acceptor vacancy V^{2-} , a donor vacancy V^+ , and so forth. We expect that the vacancy density of a given charge state (i.e., the number of vacancies per unit

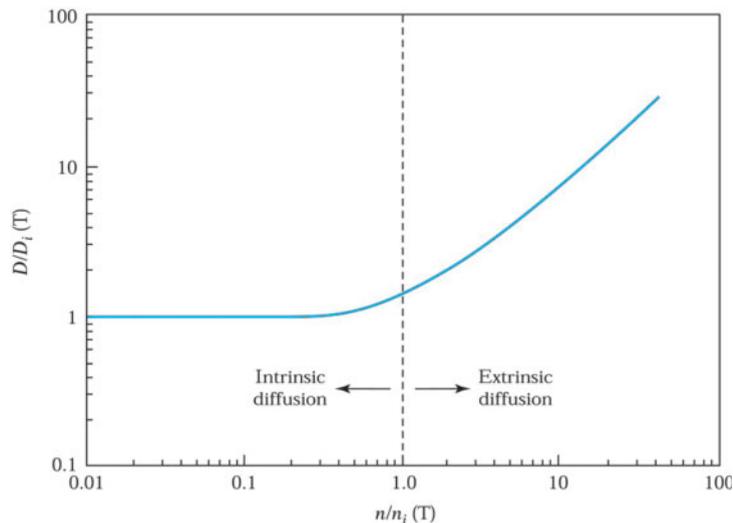


Fig. 7 Donor-impurity diffusivity versus electron concentration showing regions of intrinsic and extrinsic diffusion.¹⁰

volume, C_V) has a temperature dependence similar to that of the carrier density (see Eq. 28 in Chapter 1), that is,

$$C_V = C_i \exp\left(\frac{E_F - E_i}{kT}\right), \quad (21)$$

where C_i is the intrinsic vacancy density, E_F is the Fermi level, and E_i is the intrinsic Fermi level.

If the dopant diffusion is dominated by the vacancy mechanism, the diffusion coefficient is expected to be proportional to the vacancy density. At low doping concentrations ($n < n_i$), the Fermi level coincides with the intrinsic Fermi level ($E_F = E_i$). The vacancy density is equal to C_i and is independent of doping concentration. The diffusion coefficient, which is proportional to C_V , also is independent of doping concentration. At high concentrations ($n > n_i$), the Fermi level will move toward the conduction band edge (for donor-type vacancies), and the term $[\exp(E_F - E_i)/kT]$ becomes larger than unity. This causes C_V to increase, which in turn causes the diffusion coefficient to increase, as shown in the right side of Fig. 7.

When the diffusion coefficient varies with dopant concentration, Eq. 4 should be used as the diffusion equation instead of Eq. 5, in which D is independent of C . We consider the case where the diffusion coefficient can be written as

$$D = D_s \left(\frac{C}{C_s}\right)^\gamma, \quad (22)$$

where C_s is the surface concentration, D_s is the diffusion coefficient at the surface, C and D are the concentration and the diffusion coefficient in the bulk, and γ is a parameter to describe the concentration dependence. For such a case, we can write the diffusion equation, Eq. 4, as an ordinary differential equation and solve it numerically.

Figure 8 shows the solutions¹¹ for a constant-surface-concentration diffusion with different values of γ . For $\gamma = 0$, we have the case of constant diffusivity and the profile is the same as that shown in Fig. 5a. For $\gamma > 0$ the diffusivity decreases as the dopant concentration decreases, and increasingly steep and box-like concentration profiles result for increasing γ . Therefore, highly abrupt junctions are formed when diffusions are made into a background of an opposite impurity type due to higher impurity concentration. The abruptness of the doping profile results in a junction depth virtually independent of the background concentration. Note that the junction depth (see Fig. 8) is given by

$$\begin{aligned} x_j &= 1.6\sqrt{D_s t} \quad \text{for } D \sim C \quad (\gamma = 1), \\ x_j &= 1.1\sqrt{D_s t} \quad \text{for } D \sim C^2 \quad (\gamma = 2), \\ x_j &= 0.87\sqrt{D_s t} \quad \text{for } D \sim C^3 \quad (\gamma = 3). \end{aligned} \quad (23)$$

In the case of $\gamma = -2$, the diffusivity increases with decreasing concentration, which leads to a concave profile, as opposed to the convex profiles for other cases.

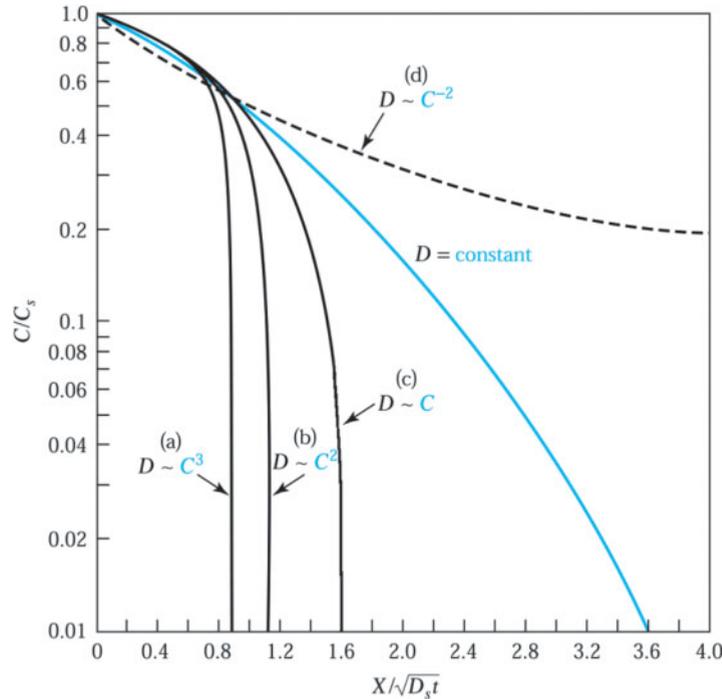


Fig. 8 Normalized diffusion profiles for extrinsic diffusion where the diffusion coefficient becomes concentration dependent.^{10,11}

14.2.2 Diffusion Profiles

Diffusion in Silicon

The measured diffusion coefficients of boron and arsenic in silicon have a concentration dependence with $\gamma \cong 1$. Their concentration profiles are abrupt, as depicted in curve *c* of Fig. 8. For gold and platinum diffusion in silicon, γ is close to -2 and their concentration profiles have the concave shape shown in curve *d* of Fig. 8.

The diffusion of phosphorus in silicon is associated with the doubly charged acceptor vacancy V^{2-} , and the diffusion coefficient at high concentration varies as C^2 . We would expect that the diffusion profile of phosphorus resembles that shown in curve *b* of Fig. 8. However, because of a *dissociation effect*, the diffusion profile exhibits anomalous behavior.

Figure 9 shows phosphorus diffusion profiles for various surface concentrations after diffusion into silicon for 1 hour at 1000°C .¹² When the surface concentration is low, corresponding to the intrinsic diffusion region, the diffusion profile is given by an erfc (curve *a*). As the concentration increases, the profile begins to deviate from the simple expression (curves *b* and *c*). At very high concentration (curve *d*), the profile near the surface is indeed similar to that in curve *b* of Fig. 8. However, at concentration n_c , a kink occurs and is followed by a rapid diffusion in the tail region. The concentration n_c corresponds to a Fermi level 0.11 eV below the conduction band. At this energy level, the coupled impurity-vacancy pair (P^+V^{2-}) dissociates to P^+ , V^- , and an electron. Thus, the dissociation generates a large number of singly charged acceptor vacancies V^- , which in turn enhances the diffusion in the tail region of the profile. The diffusivity in the tail region is over $10^{-12} \text{ cm}^2/\text{s}$, which is about two orders of magnitude larger than the intrinsic diffusivity at 1000°C . Because of its high diffusivity, phosphorus is commonly used to form deep junctions, such as the *n*-tubs in CMOS.

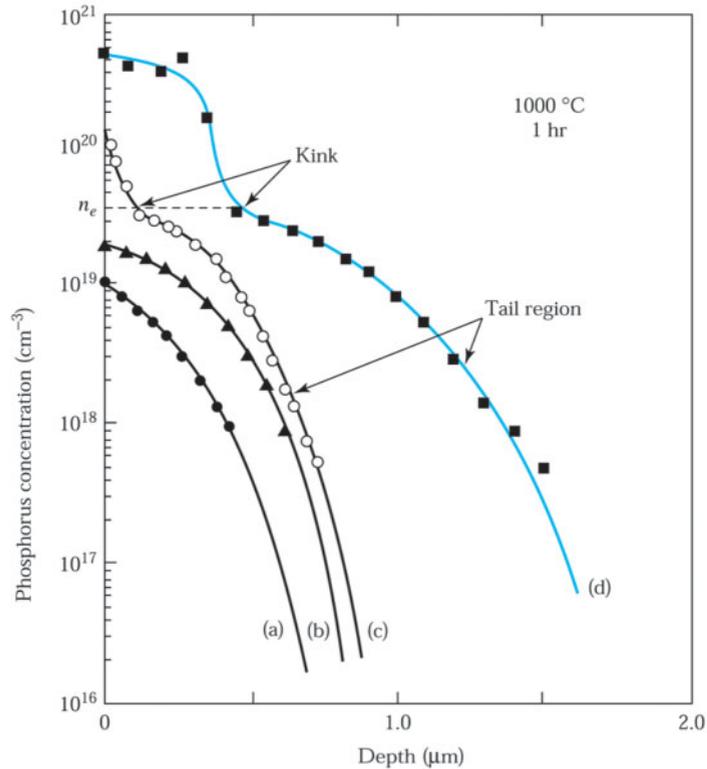


Fig. 9 Phosphorus diffusion profiles¹² for various surface concentrations after diffusion into silicon for 1 hour at 1000°C.

Zinc Diffusion in Gallium Arsenide

We expect diffusion in gallium arsenide to be more complicated than in silicon because the diffusion of impurities may involve atomic movements on both the gallium and arsenic sublattices. Vacancies play a dominant role in diffusion processes in gallium arsenide because both *p*- and *n*-type impurities must ultimately reside in lattice sites. However, the charge states of the vacancies have not been established.

Zinc is the most extensively studied diffusant in gallium arsenide. Its diffusion coefficient is found to vary as C^2 . Therefore, the diffusion profiles are steep, as shown¹³ in Fig. 10, and resemble curve *b* of Fig. 8. Note that even for the case of the lowest surface concentration, the diffusion is in the extrinsic-diffusion region, because n_i for GaAs at 1000°C is less than 10^{18} cm^{-3} . As can be seen in Fig. 10, the surface concentration has a profound effect on the junction depth. The diffusivity varies linearly with the partial pressure of the zinc vapor, and the surface concentration is proportional to the square root of the partial pressure. Therefore, from Eq. 23, the junction depth is linearly proportional to the surface concentration.

Diffusion in Strained Silicon

Strained silicon is a promising candidate for the channel of MOSFET due to the high mobility of the carriers.^{14,15} In the meantime, strain can alter the activation energy of many major steps involved in dopant diffusion, including formation of native defects and displacement of a dopant atom to form a mobile dopant complex. Strain-related bandgap narrowing can also change the charged point-defect concentration. Therefore, the diffusion associated with the defect concentration is a strong function of the strain. It impacts junction depth and effective channel length. The strain-related effects are strong for MOSFETs with scale less than 65 nm.

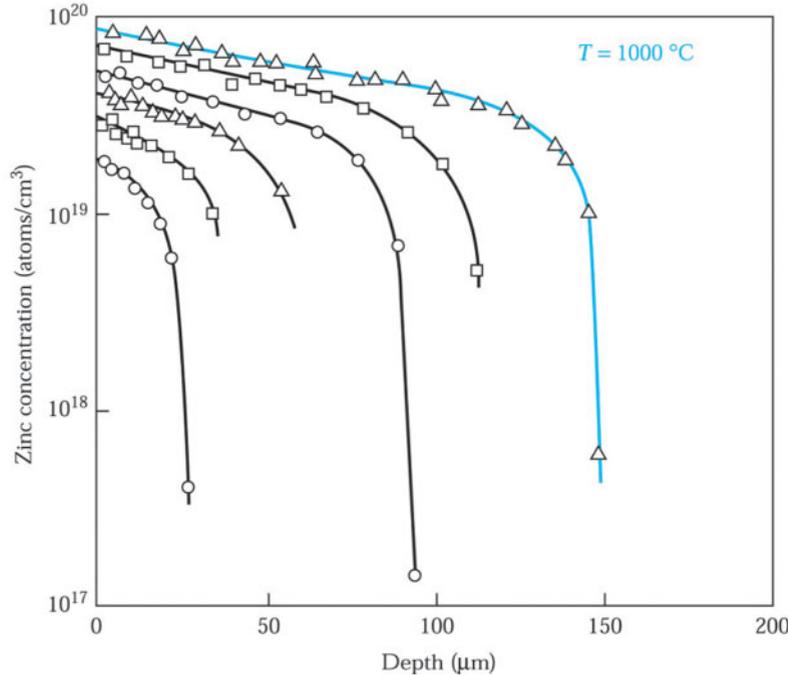


Fig. 10 Diffusion profiles¹³ of zinc in GaAs after annealing at 1000°C for 2.7 hours. The different surface concentrations are obtained by maintaining the Zn source at temperatures in the range of 600°-800°C.

Under compressive stress, the lattice constant of Si near the surface becomes smaller. To relieve it, Si atoms near the surface jump to the surface, and vacancies are created near the surface. These vacancies diffuse into the bulk and recombine with self-interstitials to decrease self-interstitials. As discussed in Section 14.1.1, since both phosphorus and boron diffuse mainly through an interstitialcy mechanism, the compressive stress results in retarded diffusion.¹⁶ Tensile stress has opposite effects to compressive stress: the lattice constant of Si becomes larger and the interstitial-mediated diffusion is enhanced. The vacancy concentration is lower due to the lattice relaxation and the vacancy-mediated diffusion is retarded.

► 14.3 DIFFUSION-RELATED PROCESSES

In this section we consider two processes in which diffusion plays an important role and the impact of these processes on device performance.

14.3.1 Lateral Diffusion

The one-dimensional diffusion equation discussed previously can describe the diffusion process satisfactorily except at the edge of the mask window. Here the impurities will diffuse downward and sideways (i.e., laterally). In this case, we must consider a two-dimensional diffusion equation and use a numerical technique to obtain the diffusion profiles under different initial and boundary conditions.

Figure 11 shows the contours of constant doping concentration for a constant-surface-concentration diffusion condition assuming that the diffusivity is independent of concentration.¹⁷ At the far right of the figure, the variation of the dopant concentration from 0.5 to $10^{-4} C/C_s$ (where C_s is the surface concentration) corresponds to the erfc distribution given by Eq. 9. The contours are in effect a map of the locations of the junctions created by diffusing into various background concentrations. For example, at $C/C_s = 10^{-4}$ (i.e., the background doping is 10^4 times lower than the surface concentration), we see from this constant-concentration curve that the vertical penetration is about 2.8 μm , whereas the lateral penetration (i.e., the penetration along the diffusion

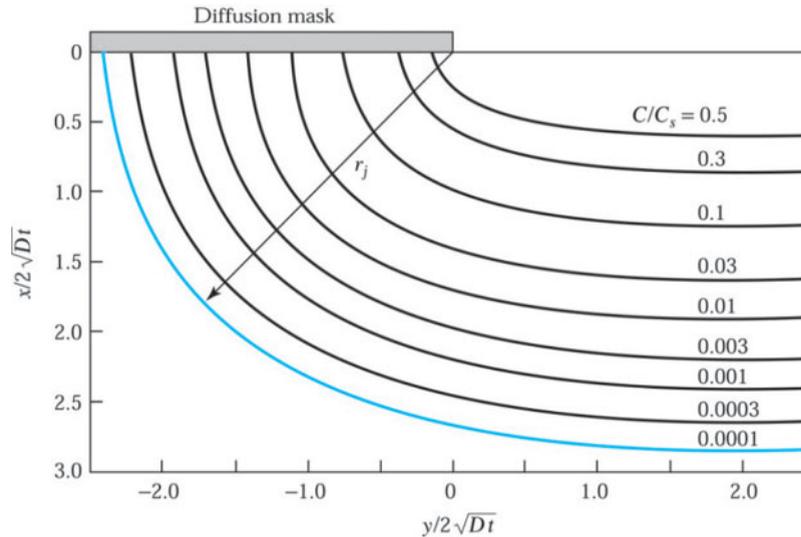


Fig. 11 Diffusion contours at the edge of an oxide window, where r_j is the radius of curvature.¹⁷

mask-semiconductor interface) is about $2.3 \mu\text{m}$. Therefore, the lateral penetration is about 80% of the penetration in the vertical direction for concentrations three or more orders of magnitude below the surface concentration. Similar results are obtained for a constant-total-dopant diffusion condition. The ratio of lateral to vertical penetration is about 75%. For concentration-dependent diffusivities, the ratio is found to be reduced slightly, to about 65%-70%.

Because of the lateral-diffusion effect, the junction consists of a central plane (or flat) region with approximately cylindrical edges with a radius of curvature r_j , as shown in Fig. 11. In addition, if the diffusion mask contains sharp corners, the shape of the junction near the corner will be roughly spherical because of lateral diffusion. Since the electric-field intensities are higher for cylindrical and spherical junction regions, the avalanche breakdown voltages of such regions can be substantially lower than that of a plane junction having the same background doping. This junction “curvature effect” was discussed in Chapter 3.

14.3.2 Impurity Redistribution During Oxidation

Dopant impurities near the silicon surface will be redistributed during thermal oxidation. The redistribution depends on several factors. When two solid phases are brought together, an impurity in one solid will redistribute between the two solids until it reaches equilibrium. This is similar to our previous discussion on impurity redistribution in crystal growth from the melt. The ratio of the equilibrium concentration of the impurity in the silicon to that in the silicon dioxide is called the *segregation coefficient* and is defined as

$$k = \frac{\text{equilibrium concentration of impurity in silicon}}{\text{equilibrium concentration of impurity in SiO}_2}. \quad (24)$$

A second factor that influences impurity distribution is that the impurity may diffuse rapidly through the silicon dioxide and escape to the gaseous ambient. If the diffusivity of the impurity in silicon dioxide is large, this factor will be important. A third factor in the redistribution process is that the oxide is growing, and thus the boundary between the silicon and the oxide advances into the silicon as a function of time. The relative rate of this advance compared with the diffusion rate of the impurity through the oxide is important in determining the extent of the redistribution. Note that even if the segregation coefficient of an impurity k equals unity, some redistribution of the impurity in the silicon will still take place. As indicated in Fig. 3 of Chapter 12, the oxide layer will be about twice as thick as the silicon layer it replaced. Therefore, the same amount of impurity will now be distributed in a larger volume, resulting in a depletion of the impurity from the silicon.

Four possible redistribution processes are illustrated⁶ in Fig. 12. These processes can be classified into two groups. In one group the oxide takes up the impurity (Fig. 12a and b for $k < 1$), and in the other the oxide rejects the impurity (Fig. 12c and d for $k > 1$). In each case, what happens depends on how rapidly the impurity can diffuse through the oxide. In group 1, the silicon surface is depleted of impurities; an example is boron with k approximately equal to 0.3. Rapid diffusion of the impurity through the silicon dioxide increases the amount of depletion; an example is boron-doped silicon heated in a hydrogen ambient because hydrogen in silicon dioxide enhances the diffusivity of boron. In group 2, k is greater than unity, so that the oxide rejects the impurity. If diffusion of the impurity through the silicon dioxide is relatively slow, the impurity piles up near the silicon surface; an example is phosphorus, with k approximately equal to 10. When diffusion through the silicon dioxide is rapid, so much impurity may escape from the solid to the gaseous ambient that the overall effect will be a depletion of the impurity; an example is gallium, with k approximately equal to 20.

The redistributed dopant impurities in silicon dioxide are seldom electrically active. However, redistribution in silicon has an important effect on processing and device performance. For example, nonuniform dopant distribution will modify the interpretation of the measurements of interface-trap properties (see Chapter 6), and a change in the surface concentration will modify the threshold voltage and device contact resistance (see Chapter 7).

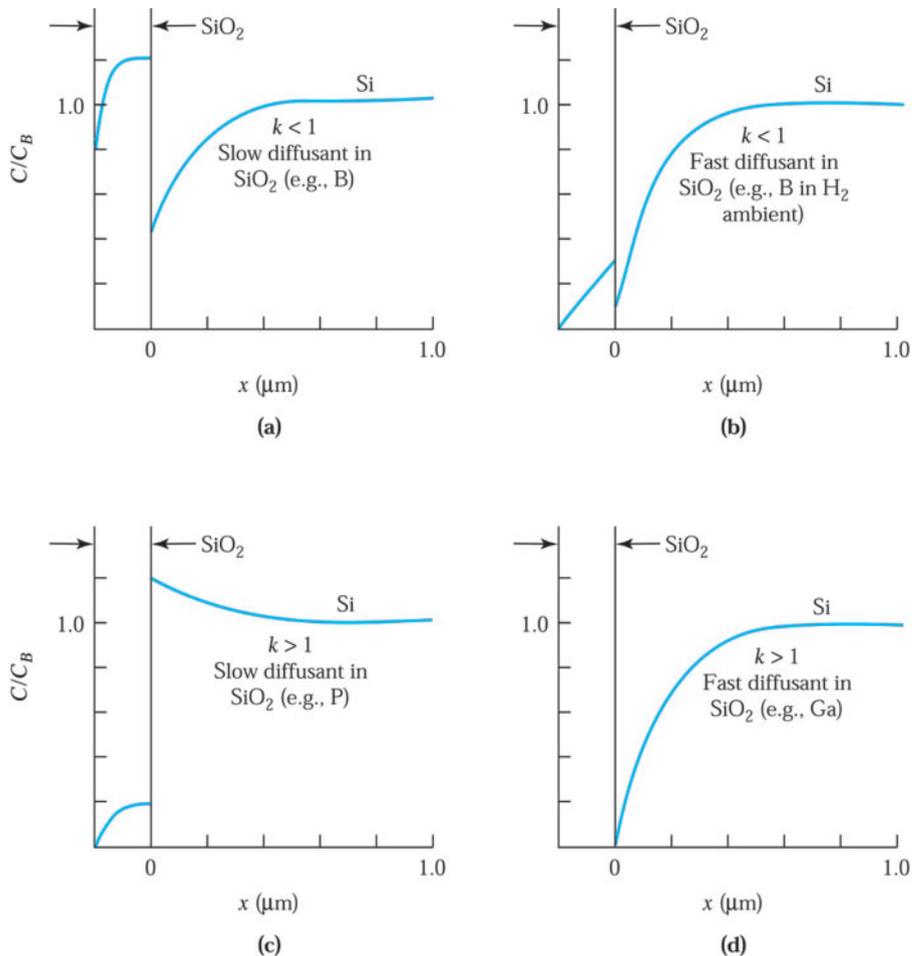


Fig. 12 Four different cases of impurity redistribution in silicon due to thermal oxidation.⁶

► 14.4 RANGE OF IMPLANTED IONS

Ion implantation is the introduction of energetic charged particles into a substrate such as silicon. Implantation energies are between 300 eV and 5 MeV, resulting in ion distributions with average depths ranging from 10 nm to 10 μm . Ion doses vary from 10^{12} ions/cm² for threshold voltage adjustment to 10^{18} ions/cm² for the formation of a buried insulating layer. Note that the dose is expressed as the number of ions implanted into 1 cm² of the semiconductor surface area. The main advantages of ion implantation are its more precise control and reproducibility of impurity dopings and its lower processing temperature than those of the diffusion process.

Basic CMOS processes usually use fifteen to seventeen ion implants per wafer. Current leading-edge CMOS processes use 20 to 23 implants, and specialized CMOS circuits (e.g., flash memory) use up to 30 implantation steps. Virtually all doping in modern CMOS devices is accomplished by ion implantation; no other technique offers comparable process control and repeatability for both the amount and position of the doping.

Figure 13 shows schematically a medium-energy ion implanter.¹⁸ The ion source has a heated filament to break up a source gas such as BF_3 or AsH_3 into charged ions (B^+ or As^+). An extraction voltage, around 40 kV, causes the charged ions to move out of the ion source chamber into a mass analyzer. The magnetic field of the analyzer is chosen such that only ions with the desired mass-to-charge ratio can travel through it without being filtered. The selected ions then enter the acceleration tube, where they are accelerated to the implantation energy as they move from high voltage to ground. Apertures ensure that the ion beam is well collimated. The pressure in the implanter is kept below 10^{-4} Pa to minimize ion scattering by gas molecules. The ion beam is then scanned over the wafer surface using electrostatic deflection plates and is implanted into the semiconductor substrate.

The energetic ions lose their energies through collision with electrons and nuclei in the substrate and finally come to rest at some depth within the lattice. The average depth can be controlled by adjusting the acceleration energy. The dopant dose can be controlled by monitoring the ion current during implantation. The principal side effect is the disruption or damage of the semiconductor lattice due to ion collisions. Therefore, a subsequent annealing treatment is needed to remove these damages.

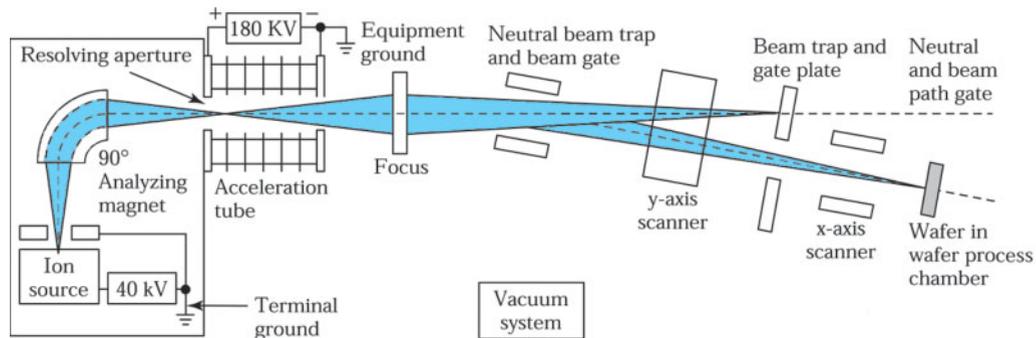


Fig. 13 Schematic of a medium-current ion implanter.

14.4.1 Ion Distribution

The total distance that an ion travels in coming to rest is called its *range* R and is illustrated¹⁹ in Fig. 14a. The projection of this distance along the axis of incidence is called the *projected range* R_p . Because the number of collisions per unit distance and the energy lost per collision are random variables, there will be a spatial distribution of ions having the same mass and the same initial energy. The statistical fluctuations in the projected range are called the *projected straggle* σ_p . There is also a statistical fluctuation along an axis perpendicular to the axis of incidence, called the *lateral straggle* σ_\perp .

Figure 14b shows the ion distribution. Along the axis of incidence, the implanted impurity profile can be approximated by a Gaussian distribution function:

$$n(x) = \frac{S}{\sqrt{2\pi}\sigma_p} \exp\left[-\frac{(x - R_p)^2}{2\sigma_p^2}\right], \tag{25}$$

where S is the ion dose per unit area. This equation is similar to Eq. 14 for constant-total-dopant diffusion, except that the quantity $4Dt$ is replaced by $2\sigma_p^2$ and the distribution is shifted along the x -axis by R_p .

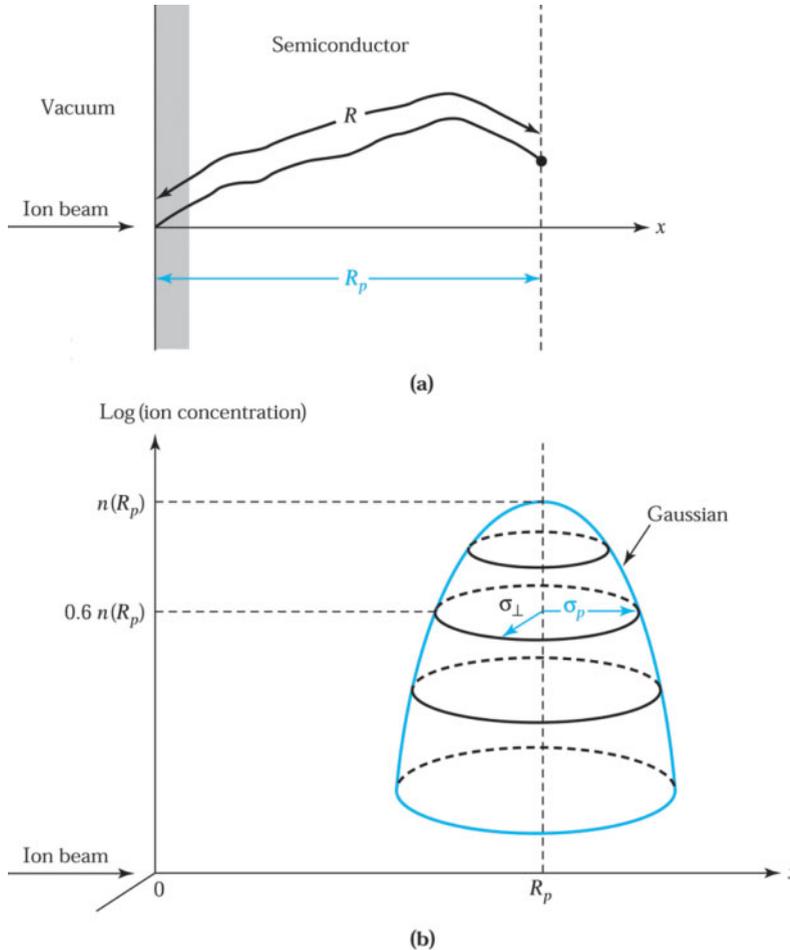


Fig. 14 (a) Schematic of the ion range R and projected range R_p . (b) Two-dimensional distribution of the implanted ions.¹⁹

Thus, for diffusion, the maximum concentration is at $x = 0$, whereas for ion implantation the maximum concentration is at the projected range R_p . The ion concentration is reduced by 40 % from its peak value at $(x - R_p) = \pm \sigma_p$, by one decade at $\pm 2\sigma_p$, by two decades at $\pm 3\sigma_p$, and by five decades at $\pm 4.8\sigma_p$.

Along the axis perpendicular to the axis of incidence, the distribution is also a Gaussian function of the form $\exp(-y^2/2\sigma_\perp^2)$. Because of this distribution, there will be some lateral implantation.²⁰ However, the lateral penetration from the mask edge (on the order of σ_\perp) is considerably smaller than that from the thermal diffusion process discussed in Section 14.3.

14.4.2 Ion Stopping

There are two stopping mechanisms by which an energetic ion, on entering a semiconductor substrate (also called the target), can be brought to rest. The first is by transferring its energy to the target nuclei. This causes deflection of the incident ion and also dislodges many target nuclei from their original lattice sites. If E is the energy of the ion at any point x along its path, we can define a nuclear stopping power $S_n(E) \equiv (dE/dx)_n$ to characterize this process. The second stopping mechanism is the interaction of the incident ion with the cloud of electrons surrounding the target's atoms. The ion loses energy in collisions with electrons through Coulombic interaction. The electrons can be excited to higher energy levels (excitation), or they can be ejected from the atom (ionization). We can define an electronic stopping power $S_e(E) \equiv (dE/dx)_e$ to characterize this process.

The average rate of energy loss with distance is given by a superposition of the above two stopping mechanisms:

$$\frac{dE}{dx} = S_n(E) + S_e(E). \quad (26)$$

If the total distance traveled by the ion before coming to rest is R , then

$$R = \int_0^R dx = \int_0^{E_0} \frac{dE}{S_n(E) + S_e(E)}, \quad (27)$$

where E_0 is the initial ion energy. The quantity R has been defined previously as the range.

We can visualize the nuclear stopping process by considering the elastic collision between an incoming hard sphere (energy E_0 and mass M_1) and a target hard sphere (initial energy zero and mass M_2),

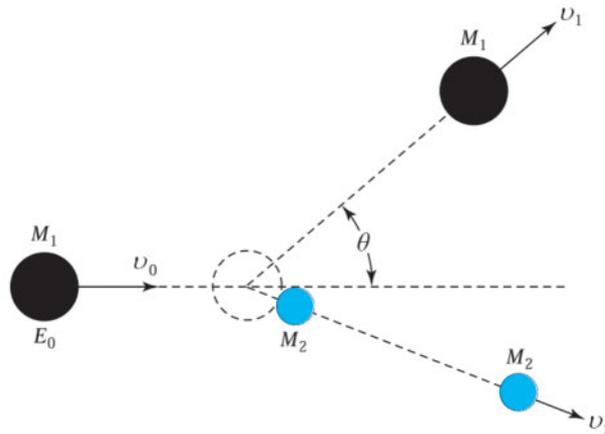


Fig. 15 Collision of hard spheres.

as illustrated in Fig. 15. When the spheres collide, momentum is transferred along the centers of the spheres. The deflection angle θ and the velocities v_1 and v_2 can be obtained from the requirements for conservation of momentum and energy. The maximum energy loss is in a head-on collision. For this case, the energy loss by the incident particle M_1 or the energy transferred to M_2 is

$$\frac{1}{2} M_2 v_2^2 = \left[\frac{4M_1 M_2}{(M_1 + M_2)^2} \right] E_0. \quad (28)$$

Since M_2 is usually of the same order of magnitude as M_1 , a large amount of energy can be transferred in the nuclear stopping process.

Detailed calculations show that the nuclear stopping power increases linearly with energy at low energies (similar to Eq. 28), and $S_n(E)$ reaches a maximum at some intermediate energy. At high energies, $S_n(E)$ becomes smaller because fast particles may not have sufficient interaction time with the target atoms to achieve effective energy transfer. The calculated values of $S_n(E)$ for arsenic, phosphorus, and boron in silicon at various energies are shown in Fig. 16 (solid line, where the superscript indicates the atomic weight).²¹ Note that heavier atoms, such as arsenic, have larger nuclear stopping power, that is, larger energy loss per unit distance.

The electronic stopping power is found to be proportional to the velocity of the incident ion, or

$$S_e(E) = k_e \sqrt{E} \quad (29)$$

where the coefficient k_e is a relatively weak function of atomic mass and atomic number. The value of k_e is approximately 10^7 (eV)^{1/2}/cm for silicon and 3×10^7 (eV)^{1/2}/cm for gallium arsenide. The electronic stopping power in silicon is plotted in Fig. 16 (dotted line). Also shown in the figure are the crossover energies at which $S_n(E)$

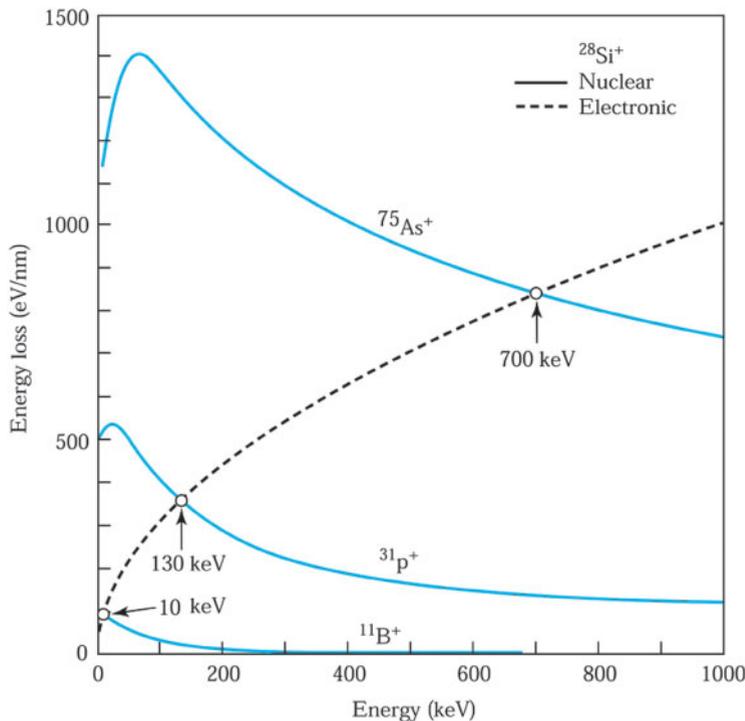


Fig. 16 Nuclear stopping power $S_n(E)$, and electronic stopping power $S_e(E)$ for As, P, and B in Si. The points of intersection of the curves correspond to the energy at which nuclear and electronic stopping are equal.²¹

equal $S_n(E)$. For boron, which has a relatively low ion mass compared with the target silicon atom, the crossover energy is only 10 keV. This means that over most of the implantation energy range of 1 keV to 1 MeV, the main energy loss mechanism is electronic stopping. On the other hand, for arsenic with relatively high ion mass, the crossover energy is 700 keV. Thus, nuclear stopping dominates over most of the energy range. For phosphorus, the crossover energy is 130 keV. For an E_0 less than 130 keV, nuclear stopping will dominate; for higher energies, electronic stopping will take over.

Once $S_n(E)$ and $S_e(E)$ are known, we can calculate the range from Eq. 27. This in turn can give us the projected range and projected straggle with the help of the following approximate equations¹⁸:

$$R_p \cong \frac{R}{1 + (M_2 / 3M_1)}, \quad (30)$$

$$\sigma_p \cong \frac{2}{3} \left[\frac{\sqrt{M_1 M_2}}{M_1 + M_2} \right] R_p. \quad (31)$$

Figure 17a shows the projected range (R_p), the projected straggle (σ_p), and the lateral straggle (σ_L) for arsenic, boron, and phosphorus in silicon.²² As expected, the larger the energy loss, the smaller the range. Also, the projected range and straggles increase with ion energy. For a given element at a specific incident energy, σ_p and σ_L are comparable and usually within $\pm 20\%$. Figure 17b shows the corresponding values for hydrogen, zinc, and

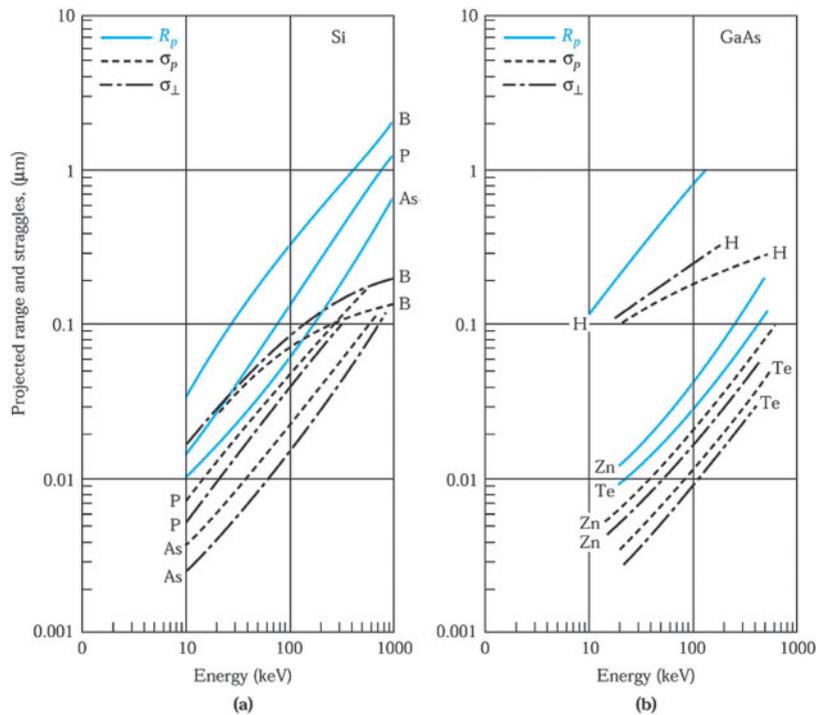


Fig. 17 Projected range, projected straggle, and lateral straggle for (a) B, P, and As in Si, and (b) H, Zn, and Te in GaAs.^{20,22}

tellurium in gallium arsenic.²⁰ Comparing Fig. 17a with Fig. 17b, we see that most of the popular dopants (except hydrogen) have larger projected ranges in silicon than they have in gallium arsenic.

► **EXAMPLE 3**

Assume 100 keV boron implants on a 200 mm silicon wafer at a dose of 5×10^{14} ions/cm². Calculate the peak concentration and the required ion beam current for 1 minute of implantation.

SOLUTION From Fig. 17a, we obtain 0.31 and 0.07 μm for the projected range and project straggle, respectively.

$$\text{From Eq. 25, } n(x) = \frac{S}{\sqrt{2\pi}\sigma_p} \exp\left[-\frac{(x-R_p)^2}{2\sigma_p^2}\right],$$

$$\frac{dn}{dx} = -\frac{S}{\sqrt{2\pi}\sigma_p} \frac{2(x-R_p)}{2\sigma_p^2} \exp\left[-\frac{(x-R_p)^2}{2\sigma_p^2}\right] = 0.$$

The peak concentration is at $x = R_p$, and $n(x) = 2.85 \times 10^{19}$ ions/cm³.

The total number of implanted ions = $Q = 5 \times 10^{14} \times \pi \times \left(\frac{20}{2}\right)^2 = 1.57 \times 10^{17}$ ions.

The required ion current = $I = \frac{qQ}{t} = \frac{1.6 \times 10^{-19} \times 1.57 \times 10^{17}}{60} = 4.19 \times 10^{-4}$ A.

$$= 0.42 \text{ mA.}$$



14.4.3 Ion Channeling

The projected range and straggle of the Gaussian distribution discussed previously give a good description of the implanted ions in amorphous or fine-grained polycrystalline substrates. Both silicon and gallium arsenide behave as if they were amorphous semiconductors, provided the ion beam is misoriented from the low-index crystallographic direction (e.g., $\langle 111 \rangle$). In this situation, the doping profile described by Eq. 25 is followed closely near the peak and extended to one or two decades below the peak value. This is illustrated¹⁹ in Fig. 18. However, even for a misorientation of 7° from the $\langle 111 \rangle$ -axis, there still is a tail that varies exponentially with distance as $\exp(-x/\lambda)$, where λ is typically on the order 0.1 μm .

The exponential tail is related to the ion-channeling effect. Channeling occurs when incident ions align with a major crystallographic direction and are guided between rows of atoms in a crystal. Figure 19 illustrates a diamond lattice viewed along a $\langle 110 \rangle$ -direction.²³ Ions implanted in the $\langle 110 \rangle$ -direction will follow trajectories that will not bring them close enough to a target atom to lose significant amounts of energy in nuclear collisions. Thus, for channeled ions, the only energy loss mechanism is electronic stopping and the range of channeled ions can be significantly larger than it would be in an amorphous target. Ion channeling is particularly critical for low-energy implant and heavy ions.

Channeling can be minimized by several techniques: a blocking amorphous surface layer, wafer misorientation, and creating a damage layer in the wafer surface. The usual blocking amorphous layer is simply a thin layer of grown silicon dioxide (Fig. 20a). The layer randomizes the direction of the ion beam so that the

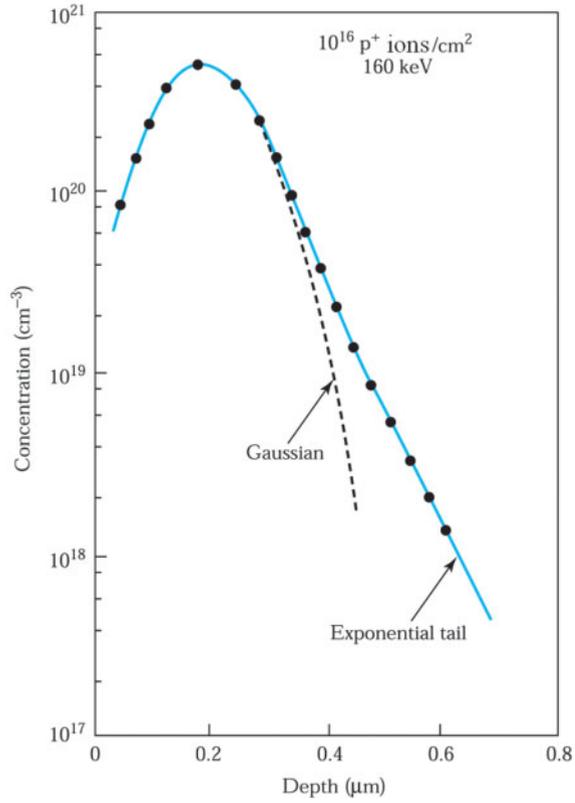


Fig. 18 Impurity profile obtained in a purposely misoriented target. Ion beam is incident 7° from the <111>-axis.¹⁹

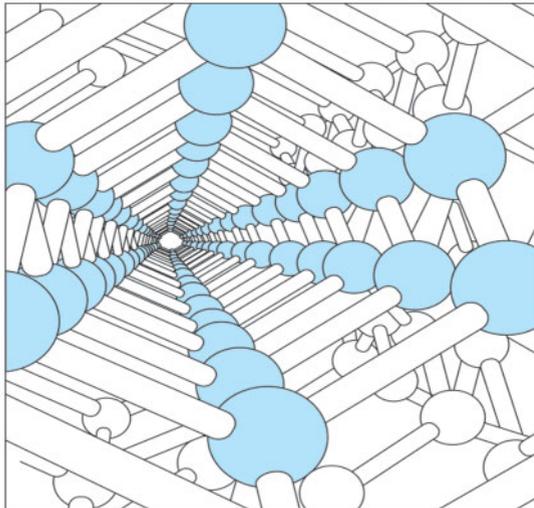


Fig. 19 Model for a diamond structure, viewed along a <110> axis.²³

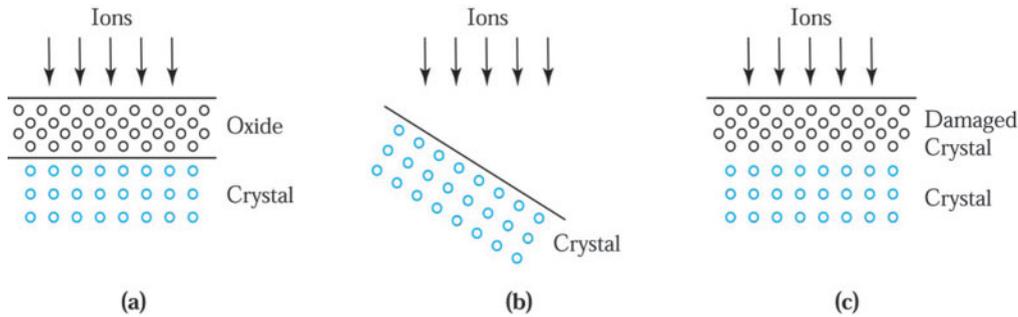


Fig. 20 (a) Implant through an amorphous oxide layer, (b) misorientation of the beam direction to all crystal axes, and (c) predamage on the crystal surface.

ions enter the wafer at different angles and not directly down the crystal channels. Misorientation of the wafers 5° - 10° from the major plane also has the effect of preventing the ions from entering the channels (Fig. 20b). With this method, most implantation machines tilt the wafer by 7° and then apply a 22° twist from the flat to prevent channeling. Predamaging the wafer surface with a heavy silicon or germanium implant creates a randomizing layer in the wafer surface (Fig. 20c). This method, however, increases the use of the expensive ion implanter and produces point defects that become leakage paths during subsequent processing.

► 14.5 IMPLANT DAMAGE AND ANNEALING

14.5.1 Implant Damage

When energetic ions enter a semiconductor substrate, they lose their energy in a series of nuclear and electronic collisions and finally come to rest. The electronic-energy loss can be accounted for in terms of electronic excitations to higher energy levels or of the generation of electron-hole pairs. However, electronic collisions do not displace semiconductor atoms from their lattice positions. Only nuclear collisions can transfer sufficient energy to the lattice that host atoms are displaced, resulting in implant damage (also called lattice disorder).²⁴ These displaced atoms may possess large fractions of the incident energy, and can in turn cause cascades of secondary displacement of nearby atoms to form a *tree of disorder* along the ion path. When the displaced atoms per unit volume approach the atomic density of the semiconductor, the material becomes amorphous.

The tree of disorder for light ions is quite different from that for heavy ions. Much of the energy loss for light ions, $^{11}\text{B}^+$ in silicon (e.g) is due to electronic collisions (see Fig. 16), which do not cause lattice damage. The ions lose their energies as they penetrate deeper into the substrate. Eventually, the ion energy is reduced below the crossover energy (10 keV for boron) where nuclear stopping becomes dominant. Therefore, most of the lattice disorder occurs near the final ion position. This is illustrated in Fig. 21a.

We can estimate the damage by considering a 100 keV boron ion. Its projected range is $0.31\ \mu\text{m}$, (Fig. 17a), and its initial nuclear energy loss is only $3\ \text{eV}/\text{\AA}$ (Fig. 16). Since the spacing between lattice planes in silicon is about $2.5\ \text{\AA}$, this means that the boron ion will lose 7.5 eV for each lattice plane because of nuclear stopping. The energy required to displace a silicon atom from its lattice position is about 15 eV. Therefore, the incident boron ion does not release enough energy from nuclear stopping to displace a silicon atom when it first enters the silicon substrate. When the ion energy is reduced to about 50 keV (at a depth of $1500\ \text{\AA}$), the energy loss due to nuclear stopping increases to 15 eV for each lattice plane (i.e., $6\ \text{eV}/\text{\AA}$), sufficient to create a lattice disorder. Assuming that one atom is displaced per lattice plane for moves roughly $25\ \text{\AA}$ from its original position, the damage volume is

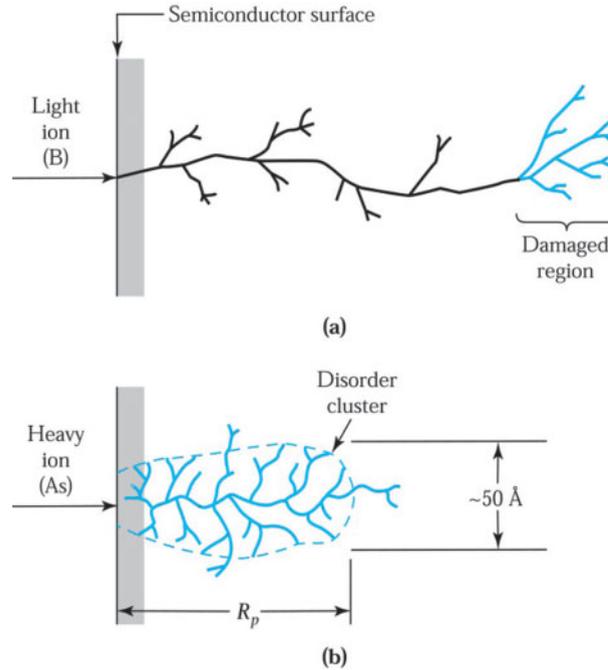


Fig. 21 Implantation disorder caused by (a) light ions and (b) heavy ions.^{2,18}

given by $V_D \cong \pi (25 \text{ \AA})^2 (1500 \text{ \AA}) = 3 \times 10^{-18} \text{ cm}^3$. The damage density is $600/V_D = 2 \times 10^{20} \text{ cm}^{-3}$, which is only 0.4% of the atoms. Thus, very high doses of light ions are needed to create an amorphous layer.

For heavy ions, the energy loss is primarily due to nuclear collisions; therefore, we expect substantial damage. Consider a 100 keV arsenic ion with a projected range of 0.06 μm or 60 nm. The average nuclear energy loss over the entire energy range is about 1320 eV/nm (Fig. 16). This means that the arsenic ion loses about 330 eV for each lattice plane on the average. Most of the energy is given to one primary silicon atom. Each primary atom will subsequently displace 22 target atoms (i.e., 330 eV/15 eV). The total number of displaced atoms is 5280. Assuming a range of 2.5 nm for the displaced atoms, the damage volume is $V_D \cong \pi (25 \text{ nm})^2 (60 \text{ nm}) = 10^{-18} \text{ cm}^3$. The damage density is then $5280/V_D \cong 5 \times 10^{21} \text{ cm}^{-3}$, or about 10% of the total number of atoms in V_D . As a result of the heavy-ion implantation, the material has become essentially amorphous. Figure 21b illustrates the situation where the damage forms a disordered cluster over the entire projected range.

To estimate the dose required to convert a crystalline material to an amorphous form, we can use the criterion that the energy density is of the same order of magnitude as that needed for melting the material (i.e., 10^{21} keV/cm^3). For 100 keV arsenic ions, the dose required to make amorphous silicon is then

$$S = \frac{(10^{21} \text{ keV/cm}^3) R_p}{E_0} = 6 \times 10^{13} \text{ ions/cm}^2. \quad (32)$$

For 100 keV boron ions, the dose required is $3 \times 10^{14} \text{ ions/cm}^2$ because R_p for boron is five times larger than for arsenic. However, in practice, higher doses ($>10^{16} \text{ ions/cm}^2$) are required for boron implantation into a target at room temperature because of the nonuniform distribution of the damage along the ion path.

14.5.2 Annealing

Because of the damaged region and the disorder cluster that result from ion implantation, semiconductor parameters such as mobility and lifetime are severely degraded. In addition, most of the ions as implanted are not located in substitutional sites. To activate the implanted ions and restore mobility and other material parameters,

we must anneal the semiconductor at an appropriate combination of time and temperature. Annealing is a heat treatment that alters the microstructure of a material, causing changes in properties.

In conventional annealing, we use an open-tube batch-furnace system similar to that for thermal oxidation. The wafers are in an isothermal environment: the furnace walls are at the same temperature as the wafers. This process requires long time and high temperature to remove the implant damages. However, conventional annealing may cause substantial dopant diffusion and cannot meet the requirement for shallow junctions and narrow doping profiles. Rapid thermal annealing (RTA) is an annealing process that employs a variety of energy sources with a wide range of times, from 100 seconds down to nanoseconds—all short compared with conventional annealing. RTA can activate the dopant fully with minimal redistribution.

Conventional Annealing of B and P

Annealing characteristics depend on the dopant type and the dose involved. Figure 22 shows the annealing behaviors of boron and phosphorus implantation into silicon substrates.²² The substrate is held at room temperature (T_s) during implantation. For a given ion dose, the annealing temperature is defined as the temperature at which 90% of the implanted ions are activated by a 30-minute annealing in a conventional annealing furnace. For boron implantation, higher annealing temperatures are needed for higher doses. For phosphorus at lower doses, the annealing behavior is similar to that for boron. However, when the dose is greater than 10^{15} cm⁻², the annealing temperature drops to about 600°C. This phenomenon is related to the solid-phase epitaxy process. At phosphorus doses greater than 6×10^{14} cm⁻² the silicon surface layer becomes amorphous. The single-crystal semiconductor underneath the amorphous layer serves as a seeding area for recrystallization of the amorphous layer. The epitaxial-growth rate along the <100> direction is 10 nm/min at 550°C and 50 nm/min at 600°C, with an activation energy at 2.4 eV. Therefore, a 100-500 nm amorphous layer can be recrystallized in a few minutes. During the solid-phase epitaxial process, the impurity dopant atoms are incorporated into the lattice sites along with the host atoms; thus, full activation can be obtained at relatively low temperatures.

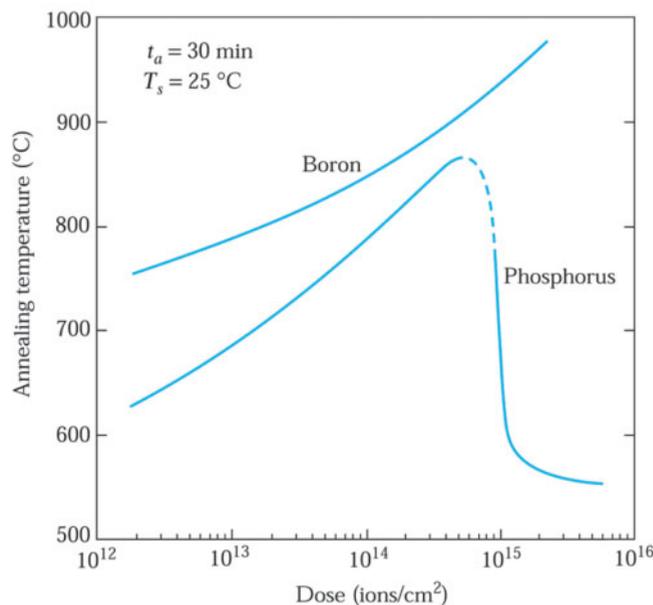


Fig. 22 Annealing temperature versus dose for 90% activation of boron and phosphorus.

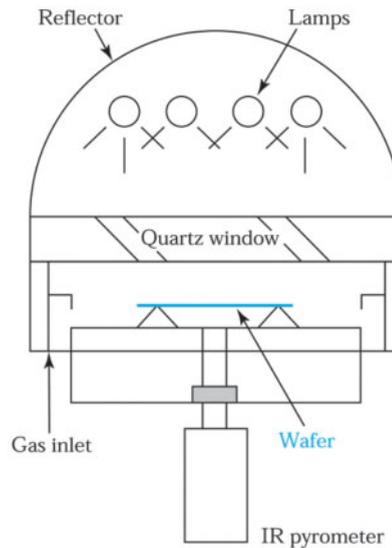


Fig. 23 Rapid thermal annealing system that is optically heated.

Rapid Thermal Annealing

A rapid thermal annealing system with transient lamp heating is shown in Fig. 23. Typical lamps in a RTA system are tungsten filaments or arc lamps. The processing chamber is made of quartz, silicon carbide, stainless steel, or aluminum, and has quartz windows through which the optical radiation passes to illuminate the wafer. The wafer holder is often made of quartz and contacts the wafer in a minimum number of places. A measurement system is placed in a control loop to set wafer temperature. The RTA system interfaces with a gas-handling system and a computer that controls system operation. Typically, wafer temperature in a RTA system is measured with a noncontact optical pyrometer that determines wafer's temperature.

In RTA, wafer is heated quickly under atmospheric conditions or at low pressure. However, the wafer is not in thermal equilibrium with its environment. The tungsten-halogen lamps (1500~2500 °C) are much hotter than the wafers (600~1100 °C)^{25,26} and the chamber walls (20~500 °C) are usually much cooler than the wafer. It is these temperature differences that permit rapid heating and cooling of the wafer. Because of the high temperatures of both the wafer and the lamps, the physics of RTA is dominated by radiation heat transfer, and the optical properties of the wafer and the chamber play an important part in the behavior.

One key advantage of RTA for ion implanted layers is the ability to reduce transient-enhanced diffusion. Transient-enhanced diffusion is the very large increase in dopant diffusivity in ion-implanted silicon, which is from the large excess of point defects that result from the ion implantation process. The phenomenon is especially severe for boron doping, since boron is already a fast diffuser, and its diffusivity is increased by silicon interstitials. Transient-enhanced diffusion effects are more severe at lower temperatures, because the degree of the excess of silicon interstitials over the equilibrium value (super-saturation) is greater at lower temperatures than higher ones. As a result, annealing at higher temperatures reduces the effects of transient-enhanced diffusion, so long as the heating cycle can be kept sufficiently short.

A comparison between conventional furnace and RTA technologies is shown in Table 2. To achieve short processing times using RTA, trade-offs must be made in temperature and process uniformity, temperature measurement and control, and wafer stress and throughput. In addition, there are concerns over the introduction of electrically active wafer defects during the very fast (100°-300°C/s) thermal transients. Rapid heating with temperature gradients in the wafers can cause wafer damage in the form of slip dislocations induced by thermal stress. On the other hand, conventional furnace processing brings with it significant problems, such as particle generation from the hot walls, limited ambient control in an open system, and a large thermal mass that restricts controlled heating times to tens of minutes. In fact, requirements on contamination, process control, and cost of manufacturing floor space have resulted in the paradigm shift to the RTA process.

Other Applications of Rapid Thermal Processing

Rapid thermal processing (RTP), which includes RTA, is a key technology in the fabrication of advanced integrated circuits, with a wide range of applications. In addition, for ion implantation damage annealing and dopant activation, RTP is also used for metal silicide and nitride formation, dielectric formation and annealing, and reflow of deposited oxides.²⁶ Typical RTP systems use radiant energy sources, often tungsten-halogen lamps, to heat a wafer to a high temperature for a period of less than a minute. Shrinking device dimensions and increasing wafer diameters are expected to make the use of RTP even more widespread as a result of its low thermal budget, fast cycle time, and compatibility with single-wafer processing. New applications, including gate dielectric formation and rapid thermal chemical vapor deposition (RTCVD), are also emerging.

TABLE 2 TECHNOLOGY COMPARISON

| Determinant | Conventional furnace | Rapid thermal annealing |
|------------------------------|----------------------|-------------------------|
| Process | Batch | Single-wafer |
| Furnace | Hot-wall | Cold-wall |
| Heating rate | Low | High |
| Cycle time | High | Low |
| Temperature monitor | Furnace | Wafer |
| Thermal budget | High | Low |
| Particle problem | Yes | Minimal |
| Uniformity and repeatability | High | Low |
| Throughput | High | Low |

Millisecond Annealing

As devices have scaled down to below 40 nm technology node, even 1 nm of diffusion is significant for ultra-shallow junctions. The requirement of much less diffusion leads to the need for millisecond-duration heating cycles during annealing. Adequate implant damage annealing and dopant activation in millisecond heating push the annealing peak temperature somewhat below the melting point of silicon.

The peak-temperature time for conventional RTP systems is usually limited by the maximum cooling rate of the wafer and the time to switch off the heating energy source. These factors typically limit the peak-temperature time to 1 second. With the use of arc lamp energy sources, which can be switched off very fast, the spike width can be reduced to ~0.3 second, but this is not short enough to meet future requirements.

For millisecond annealing, the wafer is rapidly heated with a fast ramp to an intermediate temperature and then a very short, high-energy pulse from an array of powerful water-wall flash-lamps produces a temperature jump on the whole front surface of the wafer. This condition allows extremely fast cooling through conduction of heat away from the surface into the bulk of the wafer. The nature of the heating cycle is illustrated in Fig. 24. The ability to adjust the intermediate temperature and the magnitude of the temperature jump provide flexibility in tuning the amount of diffusion relative to the amount of activation.

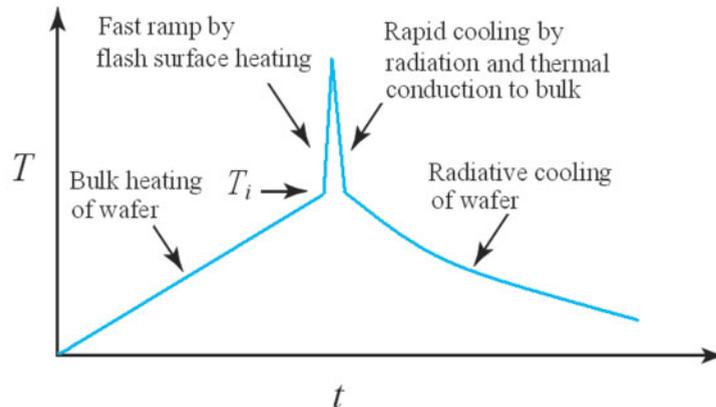


Fig. 24 The heating cycle of millisecond annealing.

► 14.6 IMPLANTATION-RELATED PROCESSES

In this section we consider a few implantation-related processes, such as multiple implantation, masking, high-angle implantation, high-energy implantation, and high-current implantation.

14.6.1 Multiple Implantation and Masking

In many applications, doping profiles other than the simple Gaussian distribution are required. One such case is the preimplantation of silicon with an inert ion to make the silicon surface region amorphous. This technique allows close control of the doping profile and permits nearly 100% dopant activation at low temperatures, as discussed previously. In such a case, a deep amorphous region may be required. To obtain this type of region, we must make a series of implants at varying ion energies and doses.

Multiple implantation can also be used to form a flat doping profile, as shown in Fig. 25. Here, four boron implants into silicon are used to provide a composite doping profile. The measured carrier concentration and that predicted using range theory are shown in the figure. Other doping profiles, unavailable from diffusion techniques, can be obtained by using various combinations of impurity dose and implantation energy. Multiple implants have been used to preserve stoichiometry during the implantation and annealing of GaAs. This approach, whereby equal amount of gallium and n -type dopant (or arsenic and p -type dopant) are implanted prior to annealing, has resulted in higher carrier activation.

To form p - n junctions in selected areas of the semiconductor substrate, an appropriate mask should be used for the implantation. Because implantation is a low-temperature process, a large variety of masking materials can be used. The minimum thickness of masking material required to stop a given percentage of incident ions can be estimated from the range parameters for ions. The inset in Fig. 26 shows the profile of an implant in a masking material. The dose implanted in the region beyond a depth d (shown shaded) is given by integration of Eq. 25 as

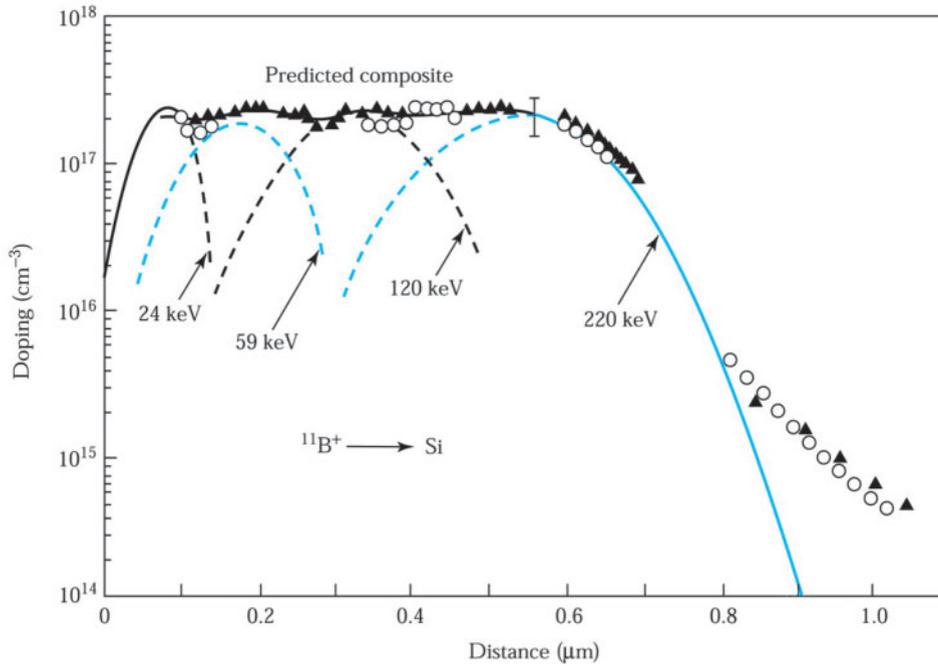


Fig. 25 Composite doping profile using multiple implants.²⁷

$$S_d = \frac{S}{\sqrt{2\pi}\sigma_p} \int_d^\infty \exp\left[-\left(\frac{x-R_p}{\sqrt{2}\sigma_p}\right)^2\right] dx. \quad (33)$$

From Table 1 we can derive the expression

$$\int_x^\infty e^{-y^2} dy = \frac{\sqrt{\pi}}{2} \operatorname{erfc}(x). \quad (34)$$

Therefore, the fraction of the dose that has “transmitted” beyond a depth d is given by the transmission coefficient T :

$$T \equiv \frac{S_d}{S} = \frac{1}{2} \operatorname{erfc}\left(\frac{d-R_p}{\sqrt{2}\sigma_p}\right). \quad (35)$$

Once T is given, we can obtain the mask thickness d from Eq. 35 for any given R_p and σ_p .

The values of d to stop 99.99% of incident ions ($T = 10^{-4}$) are shown in Fig. 26 for SiO_2 , Si_3N_4 , and a photoresist as masking materials.^{19,25} Mask thicknesses given in this figure are for boron, phosphorus, and arsenic implanted into silicon. These mask thicknesses also used as guidelines for impurity masking in gallium arsenide. The dopants are shown in parentheses. Since both R_p and σ_p vary approximately linearly with energy, the

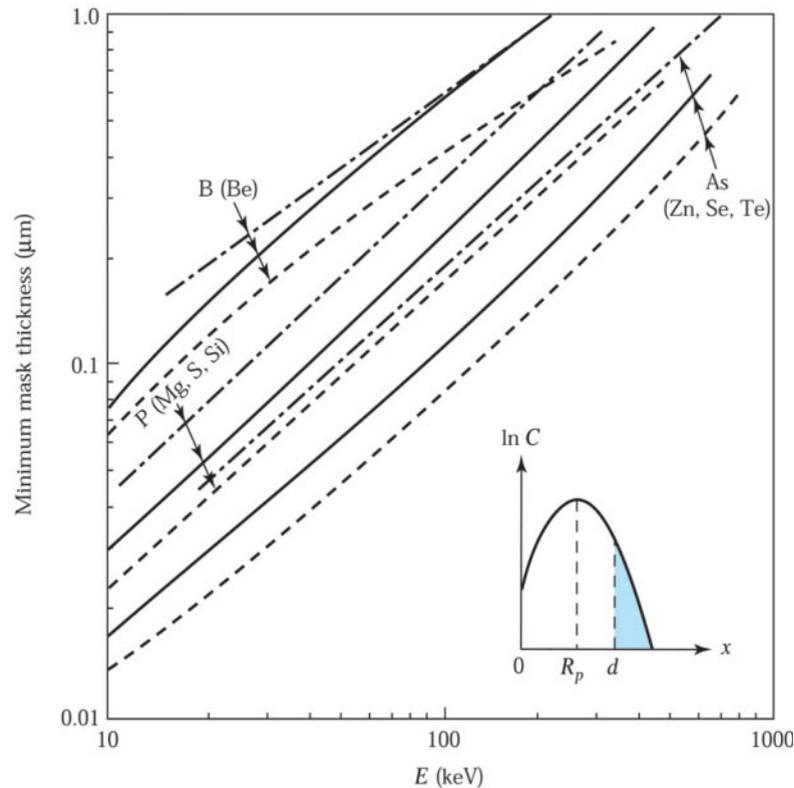


Fig. 26 Minimum thickness²⁸ of SiO_2 (—), Si_3N_4 (----), and photoresist (— · — ·) to produce a masking effectiveness of 99.99%.

minimum thickness of the masking material also increases linearly with energy. In certain applications, instead of totally stopping the beam, the masks can be used as attenuators, that can provide an amorphous surface layer to the incident ion beam to minimize the channeling effect.

► EXAMPLE 4

When boron ions are implanted at 200 keV, what thickness of SiO_2 will be required to mask 99.996% of the implanted ions ($R_p = 0.53 \mu\text{m}$, $\sigma_p = 0.093 \mu\text{m}$)?

SOLUTION The complementary error function in Eq. 35 can be approximated if the argument is large (see Table 1):

$$T \cong \frac{1}{2\sqrt{\pi}} \frac{e^{-u^2}}{u},$$

where the parameter u is given by $(d - R_p)/\sqrt{2} \sigma_p$. For $T = 10^{-4}$, we can solve the above equation to give $u = 2.8$. Thus,

$$d = R_p + 3.96\sigma_p = 0.53 + 3.96 \times 0.093 = 0.898 \mu\text{m}.$$

14.6.2 Tilt-Angle Ion Implantation

In scaling devices to submicron dimensions, it is important also to scale dopant profiles vertically. We need to produce junction depths less than 15 nm for the 28 nm technology node, including diffusion during dopant activation and subsequent processing steps. Modern device structures, such as the lightly doped drain (LDD), require precise control of dopant distributions vertically and laterally.

It is the ion velocity perpendicular to the surface that determines the projected range of an implanted ion distribution. If the wafer is tilted at a large angle to the ion beam, then the effective ion energy is greatly reduced. Figure 27 illustrates this for 60 keV arsenic ions as a function of the tilt angle, showing that it is possible to achieve extremely shallow distributions using a high tilt angle (86°). In tilt-angle ion implantation, we should consider the shadow effect (inset in Fig. 27) for the patterned wafer. A lower tilt angle leads to a small shadow area. For example, if the height of the patterned mask is $0.5\ \mu\text{m}$, with vertical sidewalls, a 7° incident ion beam will induce a 61 nm shadow. This shadow effect may introduce an unexpected series resistance in the device.

14.6.3 High-Energy and High-current Implantation

High-energy implantors, capable of energies as high as 1.5-5 MeV, are available and have been used for some novel applications. The majority of these depend on the ability to dope the semiconductor to many micrometers in depth, without the need for long diffusion times at high temperatures. High-energy implantors can also be used to produce low-resistivity buried layers. For example, a buried layer $1.5\text{-}3\ \mu\text{m}$ below the surface for a CMOS device can be achieved by high-energy implantation.

High-current implantors (10-20 mA) operating in the 25-30 keV range are routinely used for the predeposition step in diffusion technology because the amount of total dopant can be controlled precisely. After the predeposition, the dopant impurities can be driven in by a high-temperature diffusion step at the same time that

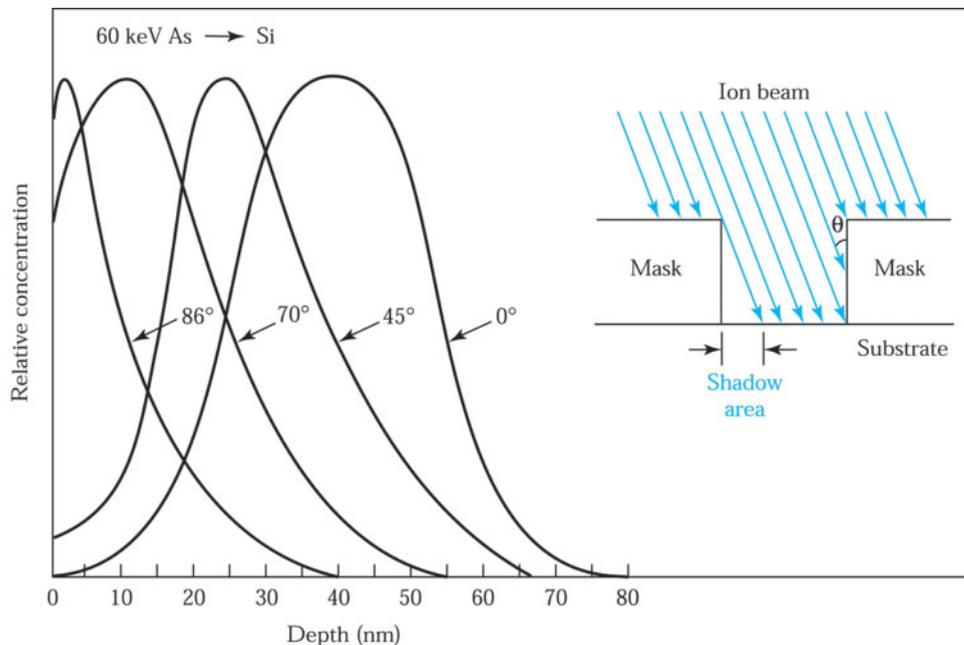


Fig. 27 60 keV arsenic implanted into silicon, as a function of beam tilt angle. Inset shows the shadow area for tilt-angle ion implantation.

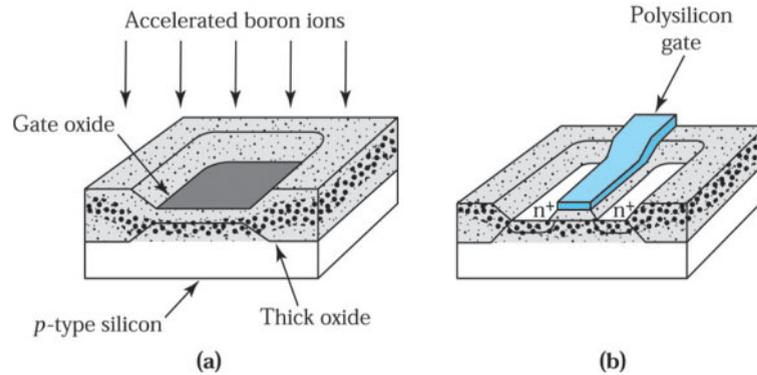


Fig. 28 Threshold voltage adjustment using boron ion implantation.²⁸

implant damage at the surface region is annealed out. Another application is the threshold voltage adjustment in MOS devices. A precisely controlled amount of dopant (e.g., boron) is implanted through the gate oxide to the channel region²⁹ (Fig. 28a). Because the projected range of boron in silicon and silicon oxide are comparable, if we choose a suitable incident energy, the ions will penetrate just the thin gate oxide, not the thicker-field oxide. The threshold voltage will vary approximately linearly with the implanted dose. After boron implantation, polysilicon can be deposited and patterned to form the gate electrode of the MOSFET. The thin oxide surrounding the gate electrode is removed, and the source and drain regions are formed as shown in Fig. 28b by another high-dose arsenic implantation.

High-current implantors with energies in the 150-200 keV range are now available. A major use for these machines is to form a buried silicon dioxide layer by implanting oxygen ions into a silicon substrate followed by a thermal annealing process. This separation by implantation of oxygen (SIMOX) is a key SOI (silicon-on-insulator) technology. The SIMOX process as shown in Fig. 29 uses a high-energy O^+ beam, typically in the 150 to 200 keV range, so that the oxygen ions have projected ranges of 100-200 nm. A heavy dose, $1-2 \times 10^{18}$ ions/cm², is used to produce an insulating layer of SiO_2 that is 100-500 nm thick. The use of SIMOX material leads to a significant reduction of source/drain capacitances in MOS devices. Moreover, it reduces coupling between devices and thus allows tighter packing without the problem of latchup. It is widely proposed as the material of choice for advanced, high-speed CMOS circuits.

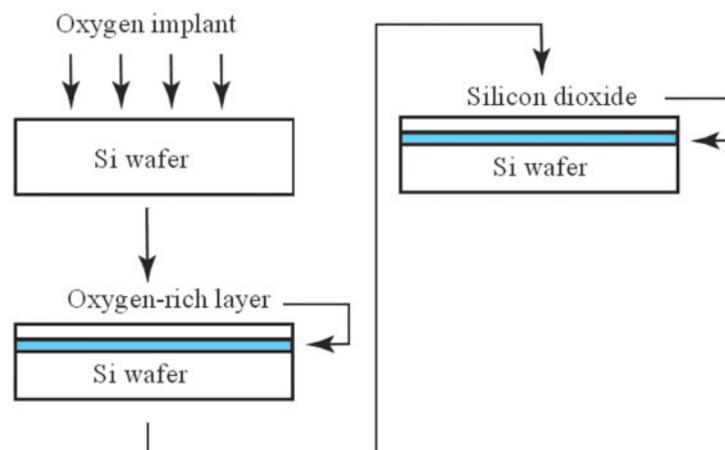


Fig. 29 The SIMOX process for SOI wafer.

Currently, 10%-20% of SOI wafers are made by the SIMOX process and 80%-90% are made by Smart Cut technology, which utilizes wafer bonding and layer transfer from a donor or seed wafer to a support handle wafer. The Smart Cut technology is shown in Fig. 30. A seed wafer is oxidized to a desired thickness. The next step is a hydrogen implantation through the oxide and into Si with a dose that is typically greater than $5 \times 10^{16} \text{ cm}^{-2}$. After the implantation, the seed wafer and the handle wafer are carefully cleaned in order to eliminate any particles and surface contaminants, and to make both surfaces hydrophilic. Wafer pairs are aligned and contacted. The wafer bonding relies on chemical assistance through hydrogen bonds and water molecules. There are a few monolayers of water trapped between two native oxide films immediately after the wafers are fused together. When the bonded wafer pair is heated to high temperature, the water will diffuse through the thin oxide to the Si interfaces to form more oxide. Finally, a complete closure of the interface occurs by coupling of the interface hydroxyl species and liberates hydrogen into the Si bulk.

The bonded wafer pairs is loaded into a furnace and heated to a temperature of $400^{\circ}\text{--}600^{\circ}\text{C}$, at which point the wafers split along the hydrogen implanted plane. The mechanism for splitting is that sufficient high-dosage hydrogen implantation will produce a sufficiently high density of platelets or microcavities. Microcavities grow to form microcracks and eventually lead to splitting of a thin layer away from the main substrate. The as-split wafer surface has a mean roughness of a few nanometers. A light touch-polish or other surface treatment brings the same surface roughness as in the standard bulk Si. The SOI technology can reduce not only the parasitic device capacitance but also the short channel effects, and thereby improve performance of scaled devices.

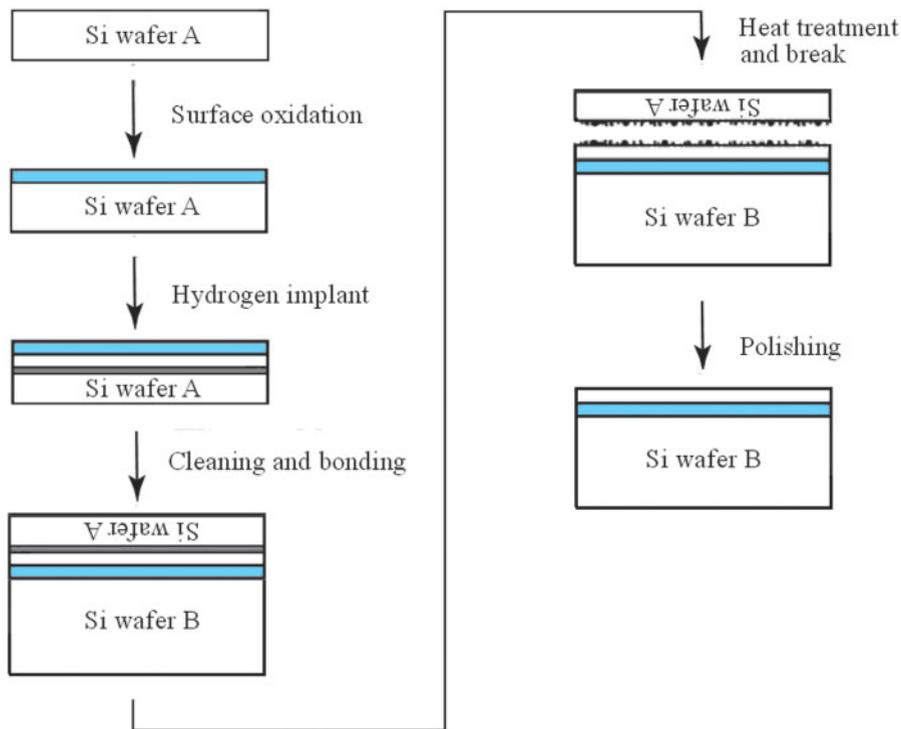


Fig. 30 The Smart Cut process for SOI wafer.

► SUMMARY

Diffusion and ion implantation are the two key methods for impurity doping. We first considered the basic diffusion equation for constant diffusivity. We obtain the complementary error function (erfc) and the Gaussian function for the constant-surface-concentration case and constant-total-dopant case, respectively. The results of a diffusion process can be evaluated by measurements of the junction depth, the sheet resistance, and the dopant profile.

When the doping concentration is higher than the intrinsic carrier concentration n_i at the diffusion temperature, the diffusivity becomes concentration dependent. This dependence has profound effects on the resulting doping profile. For example, arsenic and boron diffusivities in silicon vary linearly with the impurity concentration. Their doping profiles are much more abrupt than the erfc profile. Phosphorus diffusivity in silicon varies as the square of concentration. This dependence and a dissociation effect give rise to a phosphorus diffusivity that is 100 times larger than its intrinsic diffusivity.

Lateral diffusion at the edge of a mask and impurity redistribution during oxidation are two processes in which diffusion can have an important impact on device performance. The former can substantially reduce the breakdown voltage, and the latter will influence the threshold voltage as well as the contact resistance.

The key parameters for ion implantation are the projected range R_p and its standard deviation σ_p , also called the projected straggle. The implantation profile can be approximated by a Gaussian distribution with peak located at R_p from the surface of the semiconductor substrate. The advantages of ion implantation process are more precise control of the amount of dopant, a more reproducible doping profile, and lower processing temperature compared with the diffusion process.

We considered R_p and σ_p for various elements in silicon and gallium arsenide and discussed the channeling effect and ways to minimize this effect. However, implantation may cause severe damage to the crystal lattice. To remove the implant damage and to restore mobility and other device parameters, we must anneal the semiconductor at an appropriate combination of time and temperature. Currently, rapid thermal annealing (RTA) is preferred to conventional furnace annealing because RTA can remove implant damage without thermal broadening of the doping profile.

Ion implantation has wide applications for advanced semiconductor devices. These include (a) multiple implantation to form novel distributions, (b) selection of masking materials and thickness to stop a given percentage of incident ions from reaching the substrate, (c) tilt-angle implantation to form ultrashallow junctions, (d) high-energy implantation to form buried layers, and (e) high-current implantation for predeposition and threshold voltage adjustment and to form an insulating layer for SOI applications.

► REFERENCES

1. S. M. Sze, Ed., *VLSI Technology*, 2nd Ed., McGraw-Hill, New York, 1988, Ch. 7, 8.
2. S. K. Ghandhi, *VLSI Fabrication Principles*, 2nd Ed., Wiley, New York, 1994, Ch. 4, 6.
3. W. R. Runyan and K. E. Bean, *Semiconductor Integrated Circuit Processing Technology*, Addison-Wesley, Reading, MA, 1990, Ch. 8.
4. H. C. Casey and G. L. Pearson, "Diffusion in Semiconductors," in J. H. Crawford and L. M. Slifkin, Eds., *Point Defects in Solids*, Vol. 2, Plenum, New York, 1975.
5. J. P. Joly, "Metallic Contamination of Silicon Wafers," *Microelectron. Eng.*, **40**, 285 (1998).
6. A. S. Grove, *Physics and Technology of Semiconductor Devices*, Wiley, New York, 1967.
7. ASTM Method F374-88, "Test Method for Sheet Resistance of Silicon Epitaxial, Diffused, and Ion-implanted Layers Using a Collinear Four-Probe Array," **V10**, 249 (1993).
8. J. C. Irvin, "Evaluation of Diffused Layers in Silicon," *Bell Syst. Tech. J.*, **41**, 2 (1962).

9. ASTM Method E1438-91, "Standard Guide for Measuring Width of Interfaces in Sputter Depth Profiling Using SIMS," **V10**, 578 (1993).
10. R. B. Fair, "Concentration Profiles of Diffused Dopants," in F. F. Y. Wang, Ed., *Impurity Doping Processes in Silicon*, North-Holland, Amsterdam, 1981.
11. L. R. Weisberg and J. Blanc, "Diffusion with Interstitial-Substitutional Equilibrium, Zinc in GaAs," *Phys. Rev.*, **131**, 1548 (1963).
12. A. F. W. Willoughby, "Double-Diffusion Processes in Silicon," in F. F. Y. Wang, Ed., *Impurity Doping Processes in Silicon*, North-Holland, Amsterdam, 1981.
13. F. A. Cunnell and C. H. Gooch, "Diffusion of Zinc in Gallium Arsenide," *J. Phys. Chem. Solid*, **15**, 127 (1960).
14. M. V. Fischetti, F. Gamiz, and W. Hansch, "On the enhanced electron mobility in strained-silicon inversion layers," *J. Appl. Phys.*, **92**, 7320 (2002).
15. M. L. Lee and E. A. Fitzgerald, "Hole mobility enhancements in nanometer-scale strained-silicon heterostructures grown on Ge-rich relaxed $\text{Si}_{1-x}\text{Ge}_x$," *J. Appl. Phys.*, **94**, 2590 (2003).
16. L. Lin, T. Kirichenko, S. K. Banerjee and G. S. Hwang, "Boron Diffusion in Strained Si: A First-Principles Study," *J. Appl. Phys.*, **96**, 5543 (2004).
17. D. P. Kennedy and R. R. O'Brien, "Analysis of the Impurity Atom Distribution Near the Diffusion Mask for a Planar p-n Junction," *IBM J. Res. Dev.*, **9**, 179 (1965).
18. I. Brodie and J. J. Murray, *The Physics of Microfabrication*. Plenum, New York, 1982.
19. J. F. Gibbons, "Ion Implantation," in S. P. Keller, Ed., *Handbook on Semiconductors*, Vol. 3, North-Holland, Amsterdam, 1980.
20. S. Furukawa, H. Matsumura, and H. Ishiwara, "Theoretical Consideration on Lateral Spread of Implanted Ions," *Jpn. J. Appl. Phys.*, **11**, 134 (1972).
21. B. Smith, *Ion Implantation Range Data for Silicon and Germanium Device Technologies*, Research Studies, Forest Grove, OR, 1977.
22. K. A. Pickar, "Ion Implantation in Silicon," in R. Wolfe, Ed., *Applied Solid State Science*, Vol. 5, Academic, New York, 1975.
23. L. Pauling and R. Hayward, *The Architecture of Molecules*, Freeman, San Francisco, 1964.
24. D. K. Brice, "Recoil Contribution to Ion Implantation Energy Deposition Distribution," *J. Appl. Phys.* **46**, 3385 (1975).
25. C. Y. Chang and S. M. Sze, Eds., *VLSI Technology*, McGraw-Hill, New York, 1996, Ch. 4.
26. R. Doering and Y. Nishi, *Handbook of Semiconductor Manufacturing Technology*, 2nd Ed., CRC Press, FL, 2008.
27. D. H. Lee and J. W. Mayer, "Ion-Implanted Semiconductor Devices," *Proc. IEEE*, **62**, 1241 (1974).
28. C. Dearnaley, et al., *Ion Implantation*, North-Holland, Amsterdam, 1973.
29. W. G. Oldham, "The Fabrication of Microelectronic Circuits," in *Microelectronics*, Freeman, San Francisco, 1977.

► PROBLEMS (* DENOTES DIFFICULT PROBLEMS)

FOR SECTION 14.1 BASIC DIFFUSION PROCESS

1. Calculate the junction depth and the total amount of dopant introduced after boron predeposition performed at 950°C for 30 minutes in a neutral ambient. Assume the substrate is *n*-type silicon with $N_D = 1.8 \times 10^{16} \text{ cm}^{-3}$ and the boron surface concentration is $C_s = 1.8 \times 10^{20} \text{ cm}^{-3}$.
2. If the sample in Prob. 1 is subjected to a neutral drive-in at 1050°C for 60 minutes, calculate the diffusion profile and the junction depth.
3. Assume the measured phosphorus profile can be represented by a Gaussian function with a diffusivity $D = 2.3 \times 10^{-13} \text{ cm}^2/\text{s}$. The measured surface concentration is $1 \times 10^{18} \text{ atoms/cm}^3$ and the measured junction depth is $1 \mu\text{m}$ at a substrate concentration of 1×10^{15} . Calculate the diffusion time and the total dopant in the diffused layer.
- *4. To avoid wafer warp due to a sudden reduction in temperature, the temperature in a diffusion furnace is decreased linearly from 1000°C to 500°C in 20 minutes. What is the effective diffusion time at the initial diffusion temperature for a phosphorus diffusion in silicon?
- *5. For a low-concentration phosphorus drive-in diffusion in silicon at 1000°C, find the percentage change of surface concentration for 1% variation in diffusion time and temperature.
6. If arsenic is diffused into a thick slice of silicon doped with 10^{15} boron atoms/cm³ at a temperature of 1100°C for 3 hours, what is the final distribution of arsenic if the surface concentration is held fixed at $4 \times 10^{18} \text{ atoms/cm}^3$? What are the diffusion length and junction depth?

FOR SECTION 14.2 EXTRINSIC DIFFUSION

7. If arsenic is diffused into a thick slice of silicon doped with 10^{15} boron atoms/cm³ at a temperature of 900°C for 3 hours, what is the final distribution of arsenic if the surface concentration is held fixed at $4 \times 10^{18} \text{ atoms/cm}^3$? What is the junction depth? Assume

$$D = D_0 e^{-\frac{E_a}{kT}} \times \frac{n}{n_i}, \quad D_0 = 45.8 \text{ cm}^2/\text{s}, \quad E_a = 4.05 \text{ eV}, \quad x_j = 1.6\sqrt{Dt}.$$

8. Explain the meaning of intrinsic and extrinsic diffusion.

FOR SECTION 14.3 DIFFUSION-RELATED PROCESSES

9. Define the segregation coefficient.
10. Assume that the Cu concentration in SiO₂ layer is $5 \times 10^{13} \text{ atoms/cm}^3$ after vapor phase decomposition and is measured with atomic absorption spectrometry. The Cu concentration in the Si layer is $3 \times 10^{11} \text{ atoms/cm}^3$ after HF/H₂O₂ dissolution. Calculate the segregation coefficient of Cu in SiO₂/Si layers.

FOR SECTION 14.4 RANGE OF IMPLANTED IONS

11. In a 200 mm wafer boron-ion-implantation system, assume the beam current is 10 μA. For the *p*-channel transistor, calculate the implant time required to reduce threshold voltage from -1.1 V to -0.5 V . Assume that the implanted acceptors form a sheet of negative charge just below the Si surface and the oxide thickness is 10 nm.

12. Assume that a 100 mm diameter GaAs wafer is uniformly implanted with 100 keV zinc ions for 5 minutes with a constant ion beam current of 10 μA . What are the ion dose per unit area and the peak ion concentration?
13. A silicon p - n junction is formed by implanting boron ions at 80 keV through a window in an oxide. If the boron dose is $2 \times 10^{15} \text{ cm}^{-2}$ and the n -type substrate concentration is 10^{15} cm^{-3} , find the location of the metallurgical junction.
14. A threshold-voltage adjustment implantation is made through a 25 nm gate oxide. The substrate is a $\langle 100 \rangle$ -oriented p -type silicon with a resistivity of 10 $\Omega\text{-cm}$. If the incremental threshold voltage due to a 40 keV boron implantation is 1 V, what is the total implanted dose per unit area? Estimate the location of the peak boron concentration.
- *15. For the substrate in Prob. 14, what percentage of the total dose is in the silicon?

FOR SECTION 14.5 IMPLANT DAMAGE AND ANNEALING

16. If a 50 keV boron ion is implanted into the silicon substrate, calculate the damage density. Assume silicon atom density is $5.02 \times 10^{22} \text{ atoms/cm}^3$, the silicon displacement energy is 15 eV, the range is 2.5 nm, and the spacing between silicon lattice planes is 0.25 nm.
17. Explain why high-temperature RTA is preferable to low-temperature RTA for defect-free shallow-junction formation.
18. Estimate the implant dose required to reduce a p -channel threshold voltage by 1 V if the gate oxide is 4 nm thick. Assume that the implant voltage is adjusted so that the peak of the distribution occurs at the oxide-silicon interface. Thus, half of the implant goes into the silicon. Further, assume that 90% of the implanted ions in the silicon are electrically activated by the annealing process. These assumptions allow 45% of the implanted ions to be used for threshold adjusting. Also assume that all of the charge in the silicon is effectively at the silicon-oxide interface.

FOR SECTION 14.6 IMPLANTATION-RELATED PROCESSES

19. We would like to form 0.1 μm deep, heavily doped junctions for the source and drain regions of a submicron MOSFET. Compare the options that are available to introduce and activate dopant for this application. Which option would you recommend and why?
20. When an arsenic implant at 100 keV is used and the photoresist thickness is 400 nm, find the effectiveness of the resist mask in preventing the transmission of ions ($R_p = 0.6 \mu\text{m}$, $\sigma_p = 0.2 \mu\text{m}$). If the resist thickness is changed to 1 μm , calculate the masking efficiency.
21. With reference to Ex. 4, what thickness of SiO_2 is required to mask 99.999% of the implanted ions?

Integrated Devices

- ▶ 15.1 PASSIVE COMPONENTS
 - ▶ 15.2 BIPOLAR TECHNOLOGY
 - ▶ 15.3 MOSFET TECHNOLOGY
 - ▶ 15.4 MESFET TECHNOLOGY
 - ▶ 15.5 CHALLENGES FOR NANOELECTRONICS
 - ▶ SUMMARY
-

Microwave, photonic, and power applications generally employ discrete devices. For example, an IMPATT diode is used as a microwave generator, an injection laser as an optical source, and a thyristor as a high-power switch. However, most electronic systems are built on the integrated circuit (IC), which is an ensemble of both active (e.g., transistor) and passive devices (e.g., resistor, capacitor, and inductor) formed on and within a single-crystal semiconductor substrate and interconnected by a metallization pattern.¹ ICs have enormous advantages over discrete devices connected by wire bonding. The advantages include (a) reduction of the interconnection parasitics because an IC with multilevel metallization can substantially reduce the overall wiring length, (b) full utilization of semiconductor wafer's "real estate," because devices can be closely packed within an IC chip, and (c) drastic reduction in processing cost, because wire bonding is a time-consuming and error-prone operation.

In this chapter we combine the basic processes described in previous chapters to fabricate active and passive components in an IC. Because the key element of an IC is the transistor, specific processing sequences are developed to optimize its performance. We consider three major IC technologies associated with the three transistor families: the bipolar transistor, the MOSFET, and the MESFET.

Specifically, we cover the following topics:

- The design and fabrication of the IC resistor, capacitor, and inductor.
- The processing sequence for the standard bipolar transistor and advanced bipolar devices.
- The processing sequence for MOSFET with special emphasis on CMOS and memory devices.
- The processing sequence for high-performance MESFET and monolithic microwave ICs.
- The major challenges for future nanoelectronics, including the ultrashallow junction, ultrathin oxide, new interconnection materials, low power dissipation, and isolation.

Figure 1 illustrates the interrelationship among the major process steps used for IC fabrication. Polished wafers with a specific resistivity and orientation are used as the starting material. The film formation steps include thermally grown oxide films and deposited polysilicon, dielectric, and metal films (Chapter 12). Film formation is often followed by lithography (Chapter 13) or impurity doping (Chapter 14). Lithography is generally followed by etching, which in turn is often followed by another impurity doping or film formation. The final IC is made by sequentially transferring the patterns from each mask, level by level, onto the surface of the semiconductor wafer.

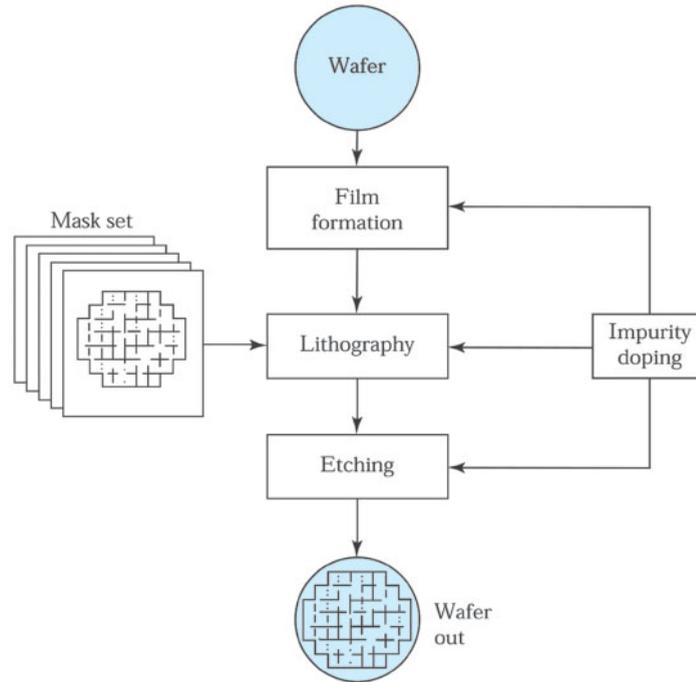


Fig. 1 Schematic flow diagram of integrated-circuit fabrication.

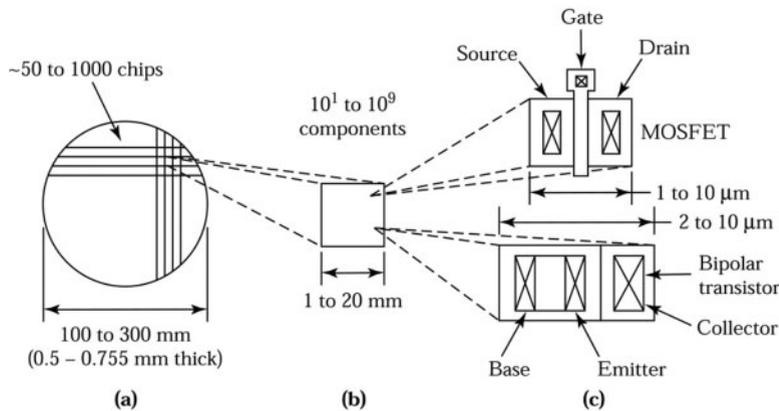


Fig. 2 Size comparison of a wafer to individual components. (a) Semiconductor wafer. (b) Chip. (c) MOSFET and bipolar transistor.

After processing, each wafer contains hundreds or thousands of identical rectangular chips (or dice), typically between 1 and 20 mm on each side, as shown in Fig. 2a. The chips are separated by sawing or laser cutting; Figure 2b shows a separated chip. Schematic top views of a single MOSFET and a single bipolar transistor are shown in Fig. 2c to give some perspective on the relative size of a component in an IC chip. Prior to chip separation, each chip is electrically tested. Defective chips are usually marked with an inkless map file. Good chips are selected and packaged to provide an appropriate thermal, electrical, and interconnection environment for electronic applications.²

IC chips may contain from a few components (transistors, diodes, resistors, capacitors, etc.) to as many as a billion or more. Since the invention of the monolithic IC in 1959, the number of components on a state-of-the-art IC chip has grown exponentially. We usually refer to the complexity of an IC as small-scale integration (SSI) for up to 100 components per chip, medium-scale integration (MSI) for up to 1000 components per chip, large-scale integration (LSI) for up to 100,000 components per chip, very-large-scale integrated (VLSI) for up to 10^7 components per chip, and ultra-large-scale integration (ULSI) for larger numbers of components per chip. In Section 15.3, we show two ULSI chips, a 48-core microprocessor chip that contains over 1.3 billion transistors (45 nm process) and an 8 Gbit dynamic random access memory (DRAM) chip that contains over 16 billion components.

► 15.1 PASSIVE COMPONENTS

15.1.1 The Integrated-Circuit Resistor

To form an IC resistor, we can deposit a resistive layer on a silicon substrate, then pattern the layer by lithography and etching. We can also define a window in a silicon dioxide layer grown thermally on a silicon substrate and then implant (or diffuse) impurities of the opposite conductivity type into the wafer. Figure 3 shows top and cross-sectional views of two resistors formed by the latter approach: one has a meander shape and the other has a bar shape.

Consider the bar-shaped resistor first. The differential conductance dG of a thin layer of the p -type material that is of thickness dx parallel to the surface and at a depth x_j (as shown by the B-B cross section) is

$$dG = q\mu_p p(x) \frac{W}{L} dx, \quad (1)$$

where W is the width of the bar, L is the length of the bar (we neglect the end contact areas for the time being), μ_p is mobility of holes, and $p(x)$ is the doping concentration. The total conductance of the entire implanted region of the bar is given by

$$G = \int_0^{x_j} dG = q \frac{W}{L} \int_0^{x_j} \mu_p p(x) dx, \quad (2)$$

where x_j is the junction depth. If the value of μ_p , which is a function of the hole concentration, and the distribution of $p(x)$ are known, the total conductance can be evaluated from Eq. 2. We can write

$$G \equiv g \frac{W}{L}, \quad (3)$$

where $g \equiv q \int_0^{x_j} \mu_p p(x) dx$ is the conductance of a square resistor pattern, that is, $G = g$ when $L = W$.

The resistance is therefore given by

$$R \equiv \frac{1}{G} = \frac{L}{W} \left(\frac{1}{g} \right), \quad (4)$$

where $1/g$ usually is defined by the symbol R_{\square} is called the sheet resistance. The sheet resistance has units of ohms but is conventionally specified in units of ohms per square (Ω/\square).

Many resistors in an integrated circuit are fabricated simultaneously by defining different geometric patterns in the mask such as those shown in Fig. 3. Since the same processing cycle is used for all these resistors, it is convenient to separate the resistance into two parts: the sheet resistance R_{\square} , determined by the implantation (or diffusion) process; and the ratio L/W , determined by the pattern dimensions. Once the value of R_{\square} is known, the

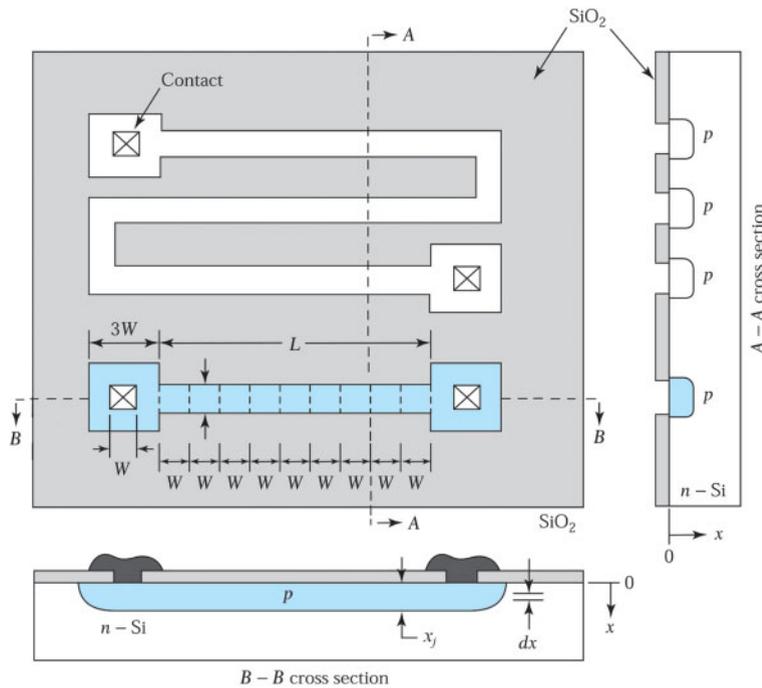


Fig. 3 Integrated-circuit resistors. All narrow lines in the large square area have the same width W , and all contacts have the same size.

resistance is given by the ratio L/W , or the number of squares (each square has an area of $W \times W$) in the resistor pattern. The end contact areas will introduce additional resistance to the IC resistors. For the type shown in Fig. 3, each end contact corresponds to approximately 0.65 square. For the meander-shape resistor, the electric-field lines at the bends are not spaced uniformly across the width of the resistor but are crowded toward the inside corner. A square at the bend does not contribute exactly 1 square but rather 0.65 square.

► EXAMPLE 1

Find the value of a resistor $90 \mu\text{m}$ long and $10 \mu\text{m}$ wide, such as the bar-shaped resistor in Fig. 3. The sheet resistance is $1 \text{ k}\Omega/\square$.

SOLUTION The resistor contains 9 squares. The two end contacts correspond to $1.3 \square$. The value of the resistor is $(9 + 1.3) \times 1 \text{ k}\Omega/\square = 10.3 \text{ k}\Omega$. ◀

15.1.2 The Integrated-Circuit Capacitor

There are basically two types of capacitors used in integrated circuits: MOS capacitors and p - n junctions. The MOS (metal-oxide-semiconductor) capacitor can be fabricated by using a heavily doped region (such as an emitter region) as one plate, the top metal electrode as the other plate, and the intervening oxide layer as the dielectric. Top and cross-sectional views of a MOS capacitor are shown in Fig. 4a. To form a MOS capacitor, a thick oxide layer is thermally grown on a silicon substrate. Next, a window is lithographically defined and then etched in the oxide. Diffusion or ion implantation is used to form a p^+ -region in the window area, while the surrounding thick oxide serves as a mask. A thin oxide layer is then thermally grown in the window area, followed by a metallization step. The capacitance per unit area is given by where ϵ_{ox} is the dielectric permittivity of silicon dioxide (the dielectric

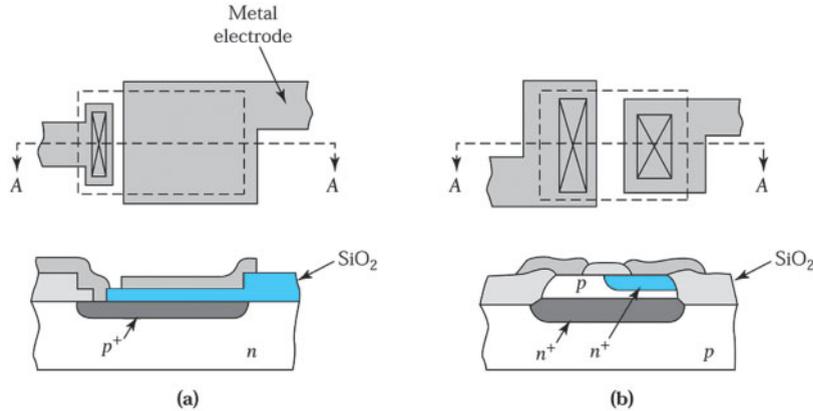


Fig. 4 (a) Integrated MOS capacitor. (b) Integrated p - n junction capacitor.

$$C = \frac{\epsilon_{ox}}{d} F / \text{cm}^2, \quad (5)$$

constant ϵ_{ox}/ϵ_0 is 3.9) and d is the thin-oxide thickness. To increase the capacitance further, insulators with higher dielectric constants are being studied, such as Si_3N_4 and Ta_2O_5 , with dielectric constants of 7 and 25, respectively. The MOS capacitance is essentially independent of the applied voltage, because the lower plate of the capacitor is made of heavily doped material. This also reduces the series resistance associated with it.

A p - n junction is sometimes used as a capacitor in an integrated circuit. The top and cross sectional views of an n^+ - p junction capacitor are shown in Fig. 4b. The detailed fabrication process is considered in Section 15.2 because this structure forms part of a bipolar transistor. As a capacitor, the device is usually reverse biased, that is, the p -region is reverse-biased with respect to the n^+ -region. The capacitance is not a constant but varies as $(V_R + V_{bi})^{-1/2}$, where V_R is the applied reverse voltage and V_{bi} is the built-in potential. The series resistance is considerably higher than that of a MOS capacitor because the p -region has higher resistivity than the p^+ -region.

► EXAMPLE 2

What are the stored charge and the number of electrons on an MOS capacitor with an area of $4 \mu\text{m}^2$ for (a) a dielectric of 10 nm thick SiO_2 and (b) a 5 nm thick Ta_2O_5 ? The applied voltage is 5 V for both cases.

SOLUTION

$$(a) \quad Q = \epsilon_{ox} \times A \times \frac{V}{d} = 3.9 \times 8.85 \times 10^{-14} \text{ F/cm} \times 4 \times 10^{-8} \text{ cm}^2 \times \frac{5 \text{ V}}{1 \times 10^{-6} \text{ cm}} = 6.9 \times 10^{-14} \text{ C}$$

or

$$Q_s = 6.9 \times 10^{-14} \text{ C}/q = 4.3 \times 10^5 \text{ electrons.}$$

(b) Changing the dielectric constant from 3.9 to 25 and the thickness from 10 nm to 5 nm, we obtain $Q = 8.85 \times 10^{-13} \text{ C}$, and $Q_s = 8.85 \times 10^{-13} \text{ C}/q = 5.53 \times 10^6$ electrons. ◀

15.1.3 The Integrated-Circuit Inductor

IC inductors have been widely used in III-V-based monolithic microwave integrated circuits (MMIC)³. With the increased speed of silicon devices and advances in multilevel interconnection technology, IC inductors have started to receive more and more attention in silicon-based radio-frequency (rf) and high-frequency applications.

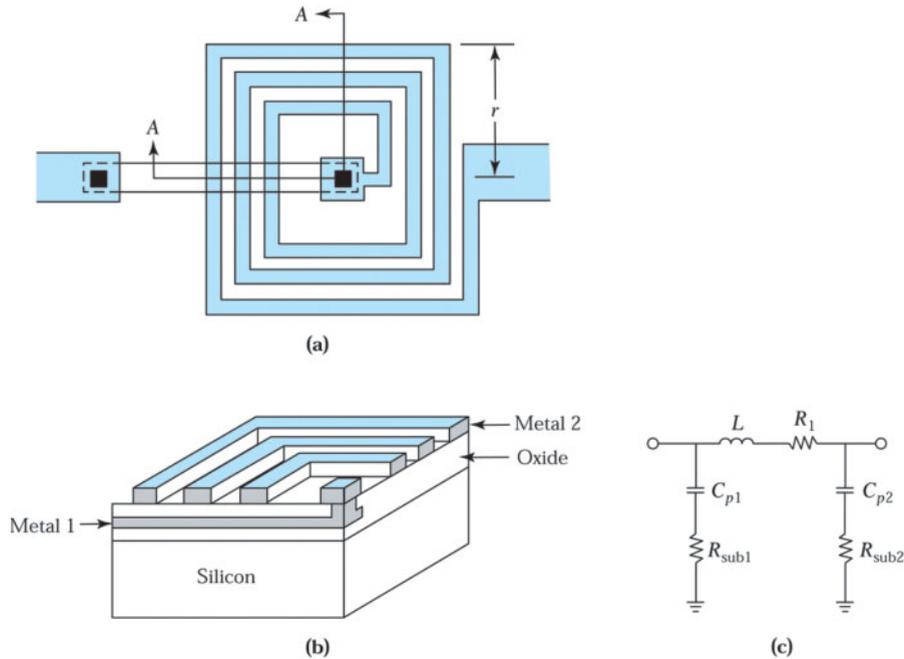


Fig. 5 (a) Schematic view of a spiral inductor on a silicon substrate. (b) Perspective view along $A-A'$. (c) An equivalent circuit model for an integrated inductor.

Many kinds of inductors can be fabricated using IC processes. The most popular method is the thin-film spiral inductor. Figure 5a and b shows the top view and cross section of a silicon-based, two-level-metal spiral inductor. To form a spiral inductor, a thick oxide is thermally grown or deposited on a silicon substrate. The first metal is then deposited and defined as one end of the inductor. Next, another dielectric is deposited onto metal 1. A via hole is defined lithographically and etched in the oxide. Metal 2 is deposited and the via hole is filled. The spiral patterned can be defined and etched on metal 2 as the second end of the inductor.

To evaluate the inductor, an important figure of merit is the quality factor, Q , defined as $Q = L\omega/R$, where L , R , and ω are the inductance, resistance, and frequency, respectively. The higher the Q values, the lower the loss from resistance, hence the better the performance of the circuits. Figure 5c shows the equivalent circuit model. R_1 is the inherent resistivity of the metal, C_{p1} and C_{p2} are the coupling capacitances between the metal lines and the substrate, and R_{sub1} and R_{sub2} are the resistances of the silicon substrate associated with the metal lines, respectively. The Q increases linearly with frequency initially and then drops at higher frequencies because of parasitic resistances and capacitances.

There are some approaches to improving the Q value. The first is to use low-dielectric-constant materials (<3.9) to reduce C_p . Another is to use a thick film metal or low-resistivity metals (e.g., Cu, Au to replace Al) to reduce the R_1 . A third approach uses an insulating substrate (e.g., silicon-on-sapphire, silicon-on-glass, or quartz) to reduce R_{sub} loss.

To obtain the exact value of a thin-film inductor, a complicated simulation tool, such as computer-aided design, must be employed for both circuit simulation and inductor optimization. The model for the thin-film inductor must take into account the resistance of the metal, the capacitance of the oxide, line-to-line capacitance, the resistance of the substrate, the capacitance of the substrate, and the inductance and mutual inductance of the metal lines.

Hence, it is more difficult to calculate the integrated inductance compared with the integrated capacitors or resistors. However, a simple equation to estimate the square planar spiral inductor is³

$$L \approx \mu_0 n^2 r \approx 1.2 \times 10^{-6} n^2 r, \quad (6)$$

where μ_0 is the permeability in vacuum ($4\pi \times 10^{-7}$ H/m), L is in henrys, n is the number of turns, and r is the radius of the spiral in meters.

► EXAMPLE 3

For an integrated inductor with an inductance of 10 nH, what is the required radius if the number of turns is 20?

SOLUTION According to Eq. 6,

$$r = \frac{10 \times 10^{-9}}{1.2 \times 10^{-6} \times 20^2} = 2.08 \times 10^{-5} \text{ (m)} = 20.8 \text{ } \mu\text{m}.$$

► 15.2 BIPOLAR TECHNOLOGY

For IC applications, especially VLSI and ULSI, the size of bipolar transistors must be reduced to meet the high-density requirement. Figure 6 illustrates the reduction in the size of the bipolar transistor in recent years.⁴ The main differences between a bipolar transistor in an IC and a discrete transistor are that all electrode contacts are located on the *top surface* of the IC wafer, and each transistor must be electrically *isolated* to prevent interactions between devices. Prior to 1970, both the lateral and vertical isolations were provided by *p-n* junctions (Fig. 6a) and the lateral *p*-isolation region was always reverse biased with respect to the *n*-type collector. In 1971, thermal oxide was used for lateral isolation, resulting in a substantial reduction in device size (Fig. 6b), because the base and collector contacts abut the isolation region. In the mid-1970s, the emitter extended to the walls of the oxide, resulting in an additional reduction in area (Fig. 6c). At the present time, all the lateral and vertical dimensions have been scaled down and emitter stripe widths have dimensions in the submicron region (Fig. 6d).

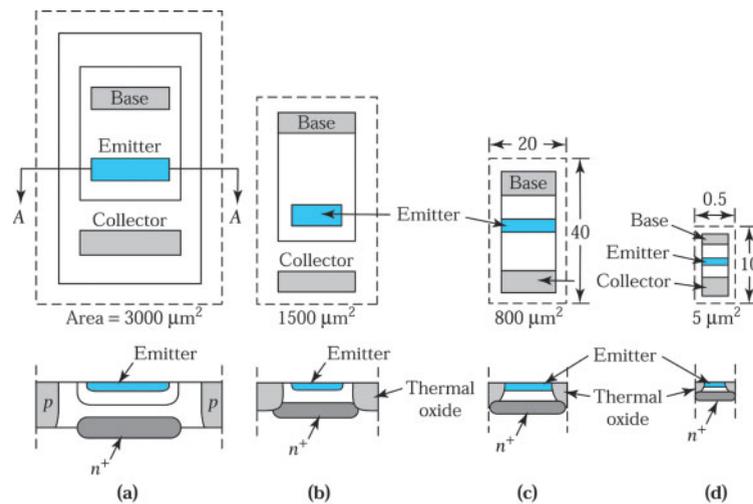


Fig. 6 Reduction of the horizontal and vertical dimensions of a bipolar transistor. (a) Junction isolation. (b) Oxide isolation. (c and d) Scaled oxide isolation.⁴

15.2.1 The Basic Fabrication Process

The majority of bipolar transistors used in ICs are of the n - p - n type because the higher mobility of minority carriers (electrons) in the base region results in higher-speed performance than can be obtained with p - n - p types. Figure 7 shows a perspective view of an n - p - n bipolar transistor, in which lateral isolation is provided by oxide walls and vertical isolation is provided by the n^+ - p junction. The lateral oxide isolation approach reduces not only the device size but also the parasitic capacitance because of the smaller dielectric constant of silicon dioxide (3.9 compared with 11.9 for silicon). We consider the major process steps that are used to fabricate the device shown in Fig. 7.

For an n - p - n bipolar transistor, the starting material is a p -type lightly doped ($\sim 10^{15} \text{ cm}^{-3}$), $\langle 111 \rangle$ - or $\langle 000 \rangle$ -oriented polished silicon wafer. Because the junctions are formed inside the semiconductor, the choice of crystal orientation is not as critical as for MOS devices. The first step is to form a buried layer. The main purpose of this layer is to minimize the series resistance of the collector. A thick oxide (0.5 – $1 \mu\text{m}$) is thermally grown on the wafer and a window is then opened in the oxide. A controlled amount of low-energy arsenic ions ($\sim 30 \text{ keV}$, $\sim 10^{15} \text{ cm}^{-2}$) is implanted into the window region to serve as a predeposit (Fig. 8a). Next, a high-temperature ($\sim 1100^\circ\text{C}$) drive-in step forms the n^+ -buried layer, which has a typical sheet resistance of $20 \Omega/\square$.

The second step is to deposit an n -type epitaxial layer. The oxide is removed and the wafer is placed in an epitaxial reactor for epitaxial growth. The thickness and the doping concentration of the epitaxial layer are determined by the ultimate use of the device. Analog circuits (with their higher voltages for amplification) require thicker layers ($\sim 10 \mu\text{m}$) and lower dopings ($\sim 5 \times 10^{15} \text{ cm}^{-3}$), whereas digital circuits (with their lower voltages for switching) require thinner layers ($\sim 3 \mu\text{m}$) and higher dopings ($\sim 2 \times 10^{16} \text{ cm}^{-3}$). Figure 8b shows a cross-sectional view of the device after the epitaxial process. Note that there is some outdiffusion from the buried layer into the epitaxial layer. To minimize the outdiffusion, a low-temperature epitaxial process should be employed and low-diffusivity impurities should be used in the buried layer (e.g., As).

The third step is to form the lateral oxide isolation region. A thin-oxide pad ($\sim 50 \text{ nm}$) is thermally grown on the epitaxial layer, followed by a silicon-nitride deposition ($\sim 100 \text{ nm}$). If nitride is deposited directly onto the silicon without the thin-oxide pad, the nitride may cause damage to the silicon surface during the subsequent high-temperature steps. Next, the nitride-oxide layers and about half of the epitaxial layer are etched using a photoresist as mask (Fig. 8c and 8d). Boron ions are then implanted into the exposed silicon areas (Fig. 8d).

The photoresist is removed and the wafer is placed in an oxidation furnace. Since the nitride layer has a very low oxidation rate, thick oxides will be grown only in the areas not protected by the nitride layer. The isolation oxide is usually grown to a thickness such that the top of the oxide becomes coplanar with the original silicon surface to minimize the surface topography. This oxide isolation process is called local oxidation of silicon (LOCOS). Figure 9a shows the cross section of the isolation oxide after the removal of the nitride layer.

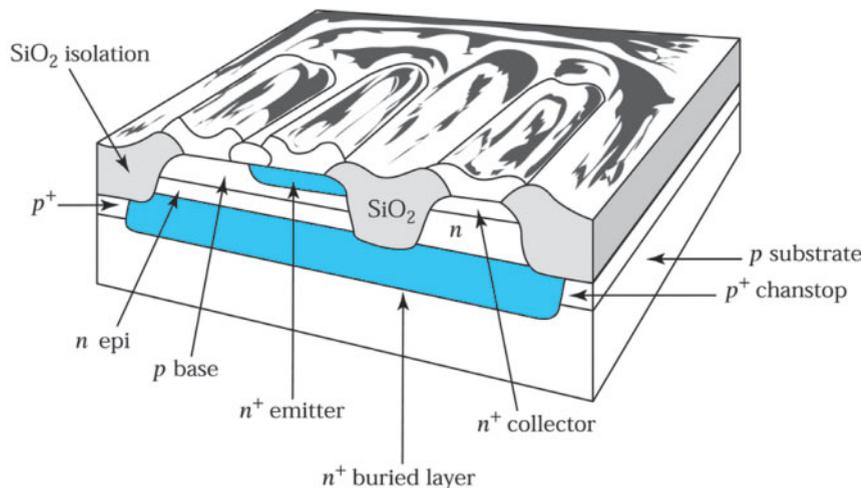


Fig. 7 Perspective view of an oxide-isolated bipolar transistor.

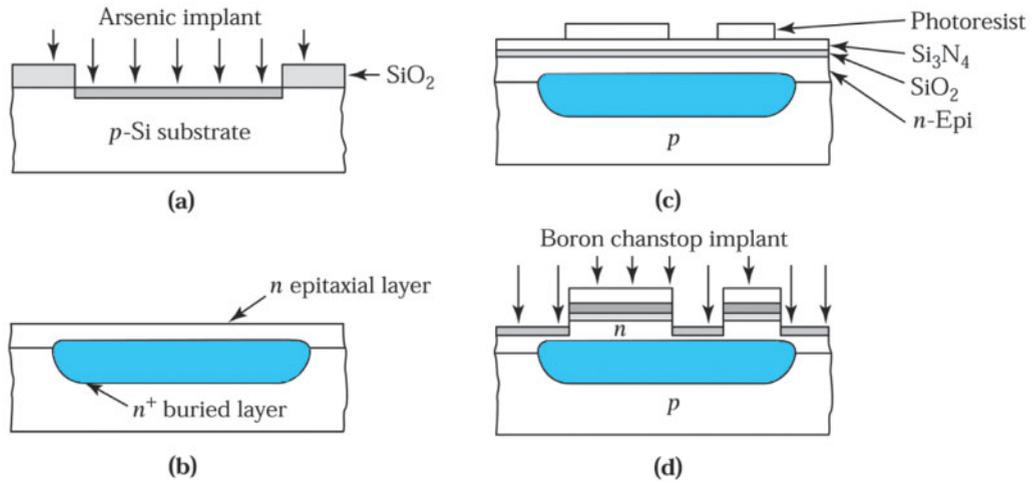


Fig. 8 Cross-sectional views of bipolar transistor fabrication. (a) Buried-layer implantation. (b) Epitaxial layer. (c) Photoresist mask. (d) Chanstop implant.

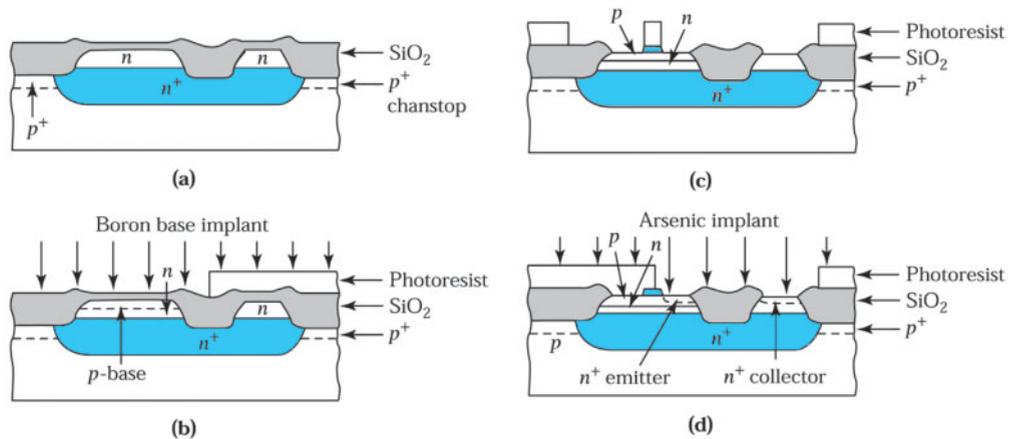


Fig. 9 Cross-section views of bipolar transistor fabrication. (a) Oxide isolation. (b) Base implant. (c) Removal of thin oxide. (d) Emitter and collector implant.

Because of segregation effects, most of the implanted boron ions are pushed underneath the isolation oxide to form a p^+ -layer. This is called a p^+ channel stop (or chanstop), because the high concentration of p -type semiconductor will prevent surface inversion and eliminate possible high-conductivity paths (or channels) among neighboring buried layers.

The fourth step is to form the base region. A photoresist is used as a mask to protect the right half of the device; then, boron ions ($\sim 10^{12} \text{ cm}^{-2}$) are implanted to form the base regions, as shown in Fig. 9b. Another lithographic process removes all the thin-pad oxide except a small area near the center of the base region (Fig. 9c).

The fifth step is to form the emitter region. As shown in Fig. 9d, the base contact area is protected by a photoresist mask; then a low-energy, high-arsenic-dose ($\sim 10^{16} \text{ cm}^{-2}$) implantation forms the n^+ -emitter and the n^+ -collector contact regions. The photoresist is removed, and a final metallization step forms the contacts to the base, emitter, and collector as shown in Fig. 7.

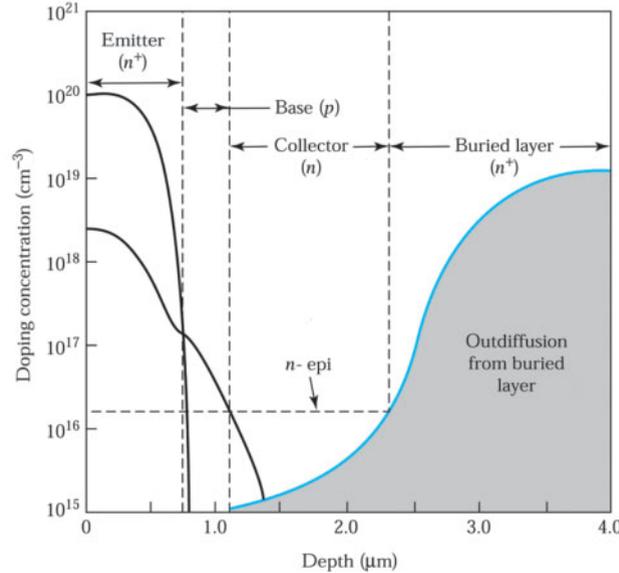


Fig. 10 *n-p-n* transistor doping profiles.

In this basic bipolar process, there are six film formation operations, six lithographic operations, four ion implantations, and four etching operations. Each operation must be precisely controlled and monitored. Failure of any one of the operations generally will render the wafer useless.

The doping profiles of the completed transistor along a coordinate perpendicular to the surface and passing through the emitter, base, and collector are shown in Fig. 10. The emitter profile is abrupt because of the concentration-dependent diffusivity of arsenic. The base doping profile beneath the emitter can be approximated by a Gaussian distribution for a limited-source diffusion. The collector doping is given by the epitaxial doping level ($\sim 2 \times 10^{16} \text{ cm}^{-3}$) for a representative switching transistor; however, at larger depths, the collector doping concentration increases because of outdiffusion from the buried layer.

15.2.2 Dielectric Isolation

In the isolation scheme described previously for the bipolar transistor, the device is isolated both from other devices by the oxide layer around its periphery and from its common substrate by a n^+p junction (buried layer). In high-voltage applications, a different approach, called dielectric isolation, is used to form insulating tubs to isolate a number of pockets of single-crystal semiconductors. In this approach the device is isolated from both its common substrate and its surrounding neighbors by a dielectric layer.

In the process for the dielectric isolation, an oxide layer is formed inside a $\langle 100 \rangle$ -oriented *n*-type silicon substrate by SIMOX process or Smart Cut technology, as discussed in Chapter 14, Section 14.6.3. Since the top silicon is so thin, the isolation region is easily formed by the LOCOS process illustrated in Fig. 8c or by etching a trench and refilling it with oxide. The other processes are almost the same as those in Figs. 8c through 9 to form the *p*-type base, n^+ -emitter, and collector.

The main advantage of this technique is its high breakdown voltage between the emitter and the collector, which can be in excess of several hundred volts. This technique is also compatible with modern CMOS integration. This CMOS-compatible process is very useful in mixed high-voltage and high-density IC.

15.2.3 Self-Aligned Double-Polysilicon Bipolar Structure

The process shown in Fig. 9c needs another lithographic process to define an oxide region to separate the base and emitter contact regions. This gives rise to a large inactive device area within the isolated boundary that increases not only the parasitic capacitances but also the resistance that degrades transistor performance. The most effective way to reduce these effects is by using the self-aligned structure.

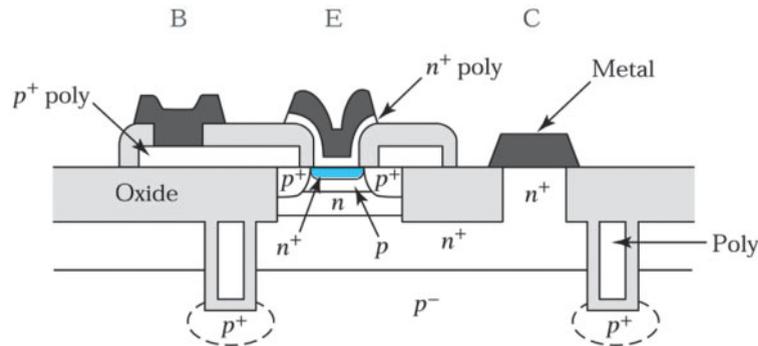


Fig. 11 Cross-section of a self-aligned, double-polysilicon bipolar transistor with advanced trench isolation.⁵

The most widely used self-aligned structure is the double-polysilicon structure with the advanced isolation provided by a trench refilled with polysilicon,⁵ as shown in Fig. 11. Figure 12 shows the detailed sequence of steps for the self-aligned double-polysilicon (n - p - n) bipolar structure.⁶ The transistor is built on an n -type epitaxial layer. A trench 5.0 μm deep is etched by reactive ion etching through the n^+ -subcollector region into the p^- substrate region. A thin layer of thermal oxide is then grown and serves as the screen oxide for the channel stop implant of boron at the bottom of the trench. The trench is then filled with undoped polysilicon and capped by a thick planar field oxide.

The first polysilicon layer is deposited and heavily doped with boron. The p^+ -polysilicon (called poly 1) will be used as a solid-phase diffusion source to form the extrinsic base region and the base electrode. This layer is covered with a chemical-vapor deposition (CVD) oxide and nitride (Fig. 12a). The emitter mask is used to pattern the emitter-area regions, and a dry-etch process is used to produce an opening in the CVD oxide and poly 1 (Fig. 12b). A thermal oxide is then grown over the etched structure, and a relatively thick oxide (approximately 0.1–0.4 μm) is grown on the vertical sidewalls of the heavily doped poly. The thickness of this oxide determines the spacing between the edges of the base and emitter contacts. The extrinsic p^+ base regions are also formed during the thermal-oxide growth step as a result of the outdiffusion of boron from the poly 1 into the substrate (Fig. 12c). Because boron diffuses laterally as well as vertically, the extrinsic base region will be able to make contact with the intrinsic base region that is formed next, under the emitter contact.

Following the oxide-grown step, the intrinsic base region is formed using ion implantation of boron (Fig. 12d). This serves to self-align the intrinsic and extrinsic base regions. After the contact is cleaned to remove any oxide layer, the second polysilicon layer is deposited and implanted with As or P. The n^+ -polysilicon (poly 2) is used as a solid-phase diffusion source to form the emitter region and the emitter electrode. A shallow emitter region is then formed through dopant outdiffusion from poly 2. A rapid thermal anneal for the base and emitter outdiffusion steps facilitates the formation of shallow emitter-base and collector-base junctions. Finally, Pt film is deposited and sintered to form PtSi over the n^+ -polysilicon emitter and the p^+ -polysilicon base contact (Fig. 12e).

This self-aligned structure allows the fabrication of emitter regions smaller than the minimum lithographic dimension. When the sidewall-spacer oxide is grown, it fills the contact hole to some extent because the thermal oxide occupies a larger volume than the original volume of polysilicon. Thus, an opening 0.8 μm wide will shrink to about 0.4 μm if sidewall oxide a 0.2 μm thick is grown on each side.

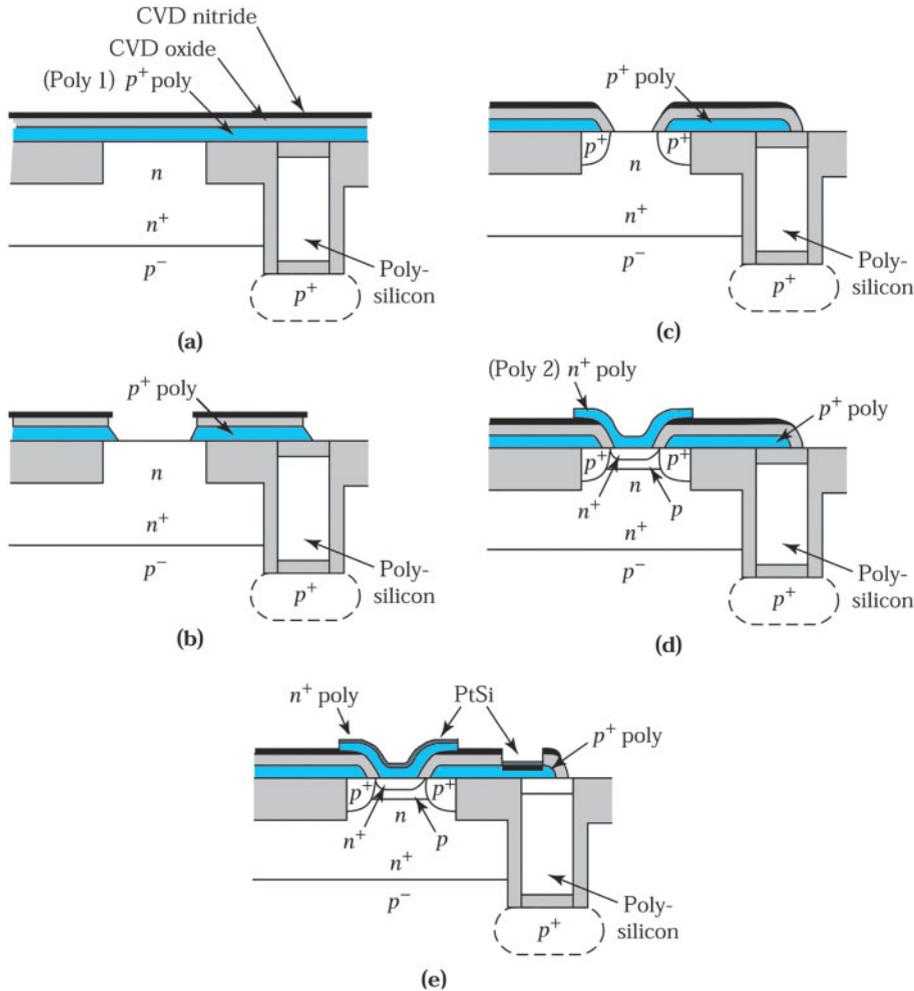


Fig. 12 Process sequence for fabricating double-polysilicon self-aligned n - p - n transistors.⁶

► 15.3 MOSFET TECHNOLOGY

At present, MOSFET is the dominant device used in ULSI circuits because it can be scaled to smaller dimensions than other types of devices. The dominant technology for MOSFET is CMOS (complementary MOSFET) technology, in which both n -channel and p -channel MOSFETs (NMOS and PMOS, respectively) are provided on the same chip. CMOS technology is particularly attractive for ULSI circuits because it has the lowest power consumption of all IC technology.

Figure 13 shows the reduction in size of the MOSFET in recent years. In the early 1970s, the gate length was $7.5\ \mu\text{m}$ and the corresponding device area was about $6000\ \mu\text{m}^2$. As the device is scaled down, there is a drastic reduction in the device area. For a MOSFET with a gate length of $0.5\ \mu\text{m}$, the device area shrinks to less than 1% of the early MOSFET. We expect that device miniaturization will continue and gate length may approach the $10\sim 20\ \text{nm}$ range by the year 2020. We consider future trends in the devices in Section 15.5.

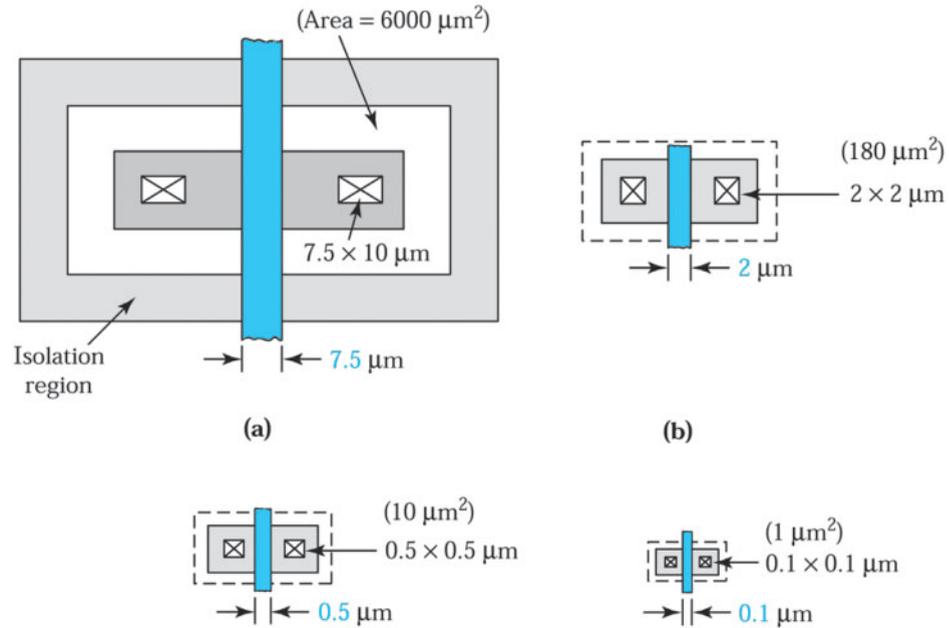


Fig. 13 Reduction in the area of the MOSFET as gate length (minimum feature length) is reduced.

15.3.1 The Basic Fabrication Process

Figure 14 shows a perspective view of an n -channel MOSFET prior to its final metallization.⁷ The top layer is a phosphorus-doped silicon dioxide (P-glass) that is used as an insulator between the polysilicon gate and the gate metallization and also as a gettering layer for mobile ions. Compare Fig. 14 with Fig. 7 for the bipolar transistor and note that a MOSFET is considerably simpler in its basic structure. Although both devices use lateral oxide isolation, there is no need for vertical isolation in the MOSFET, whereas a buried-layer n^+ - p junction is required in the bipolar transistor. The doping profile in a MOSFET is not as complicated as that in a bipolar transistor and the control of the dopant distribution is also less critical. We consider the major process steps that are used to fabricate the device shown in Fig. 14.

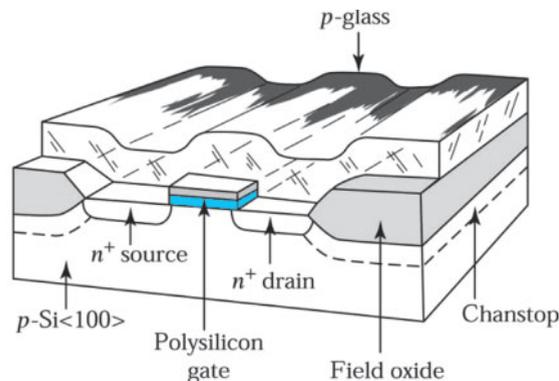


Fig. 14 Perspective view of an n -channel MOSFET.⁷

To process an n -channel MOSFET (NMOS), the starting material is a p -type, lightly doped ($\sim 10^{15} \text{ cm}^{-3}$), $\langle 100 \rangle$ -oriented polished silicon wafer. The $\langle 100 \rangle$ -orientation is preferred over $\langle 111 \rangle$ because it has an interface-trap density that is about one-tenth that of $\langle 111 \rangle$. The first step is to form the oxide isolation region using LOCOS technology. The process sequence for this step is similar to that for the bipolar transistor. A thin-pad oxide ($\sim 35 \text{ nm}$) is thermally grown, followed by a silicon nitride ($\sim 150 \text{ nm}$) deposition (Fig. 15a).⁷

The active device area is defined by a photoresist mask and a boron chanstop layer is then implanted through the composite nitride-oxide layer (Fig. 15b). The nitride layer not covered by the photoresist mask is subsequently removed by etching. After stripping the photoresist, the wafer is placed in an oxidation furnace to grow an oxide (called the field oxide), where the nitride layer is removed, and to drive in the boron implant. The thickness of the field oxide is typically $0.5\text{--}1 \mu\text{m}$.

The second step is to grow the gate oxide and to adjust the threshold voltage (see Chapter 5, Section 5.5.3). The composite nitride-oxide layer over the active device area is removed, and a thin-gate oxide layer (less than 10 nm) is grown. For an enhancement-mode n -channel device, boron ions are implanted in the channel region, as shown in Fig. 15c, to increase the threshold voltage to a predetermined value (e.g., $+0.5\text{V}$). For a depletion-mode n -channel device, arsenic ions are implanted in the channel region to decrease the threshold voltage (e.g., -0.5V).

The third step is to form the gate. A polysilicon is deposited and is heavily doped by diffusion or implantation of phosphorus to a typical sheet resistance of $20\text{--}30 \Omega/\square$. This resistance is adequate for MOSFETs with gate lengths larger than $3 \mu\text{m}$. For smaller devices, polycide, a composite layer of metal silicide and polysilicon, such as W-polycide, can be used as the gate material to reduce the sheet resistance to about $1 \Omega/\square$.

The fourth step is to form the source and drain. After the gate is patterned (Fig. 15d), it serves as a mask for the arsenic implantation ($\sim 30 \text{ keV}$, $\sim 5 \times 10^{15} \text{ cm}^{-2}$) to form the source and drain (Fig. 16a), which are self-aligned with respect to the gate.⁷ At this stage, the only overlapping of the gate is due to lateral straggling of the implanted ions (for 30 keV As , σ_{\perp} is only 5 nm). If low-temperature processes are used for subsequent steps to minimize lateral diffusion, the parasitic gate-drain and gate-source coupling capacitances can be much smaller than the gate-channel capacitance.

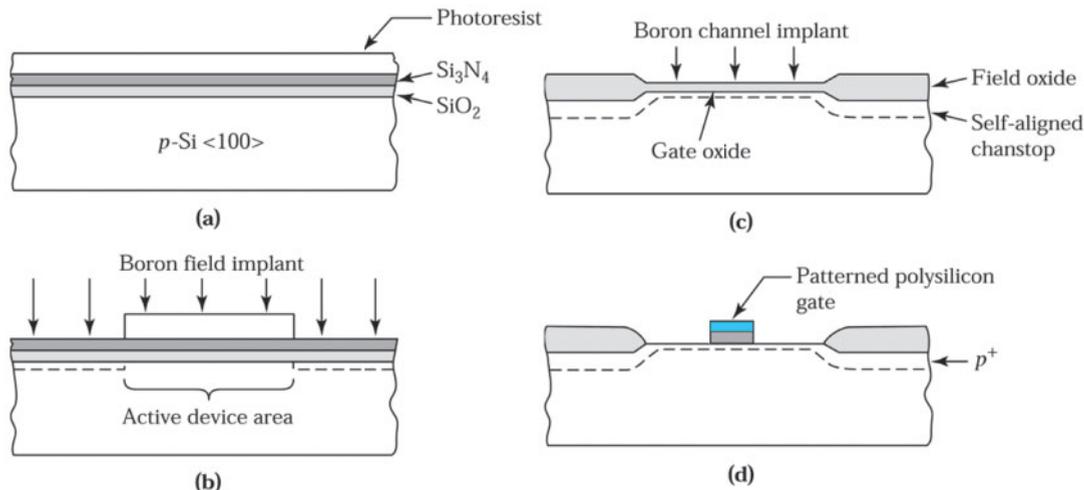


Fig. 15 Cross-sectional view of NMOS fabrication sequence.⁷ (a) Formation of SiO_2 , Si_3N_4 , and photoresist layer. (b) Boron implant. (c) Field oxide. (d) Gate.

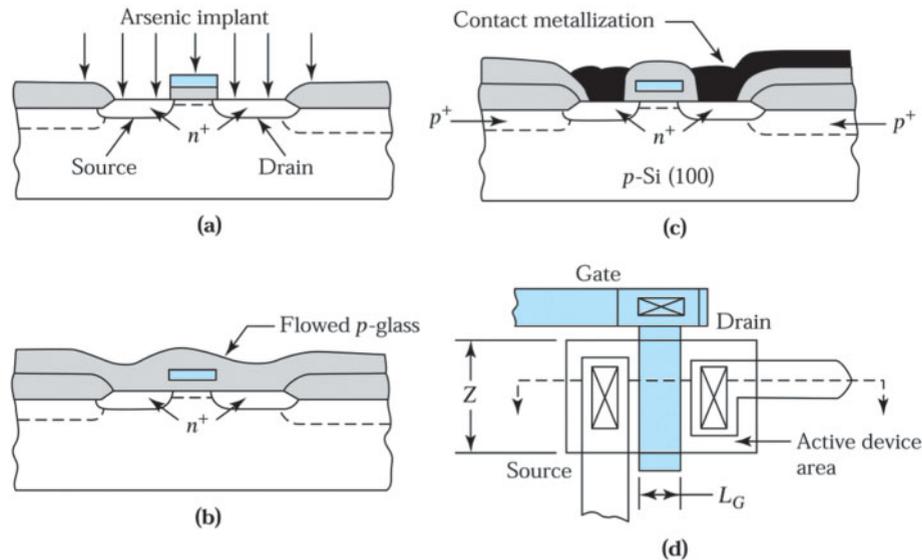


Fig. 16 NMOS fabrication sequence.⁷ (a) Source and drain. (b) P-glass deposition. (c) Cross section of the MOSFET. (d) Top view of the MOSFET.

The last step is the metallization. A phosphorus-doped oxide (P-glass) is deposited over the entire wafer and is flowed by heating the wafer to give a smooth surface topography (Fig. 16b). Contact windows are defined and etched in the P-glass. A metal layer, such as aluminum, is then deposited and patterned. A cross-section view of the completed MOSFET is shown in Fig. 16c, and the corresponding top view is shown in Fig. 16d. The gate contact is usually made outside the active device area to avoid possible damage to the thin-gate oxide.

► EXAMPLE 4

What is the maximum gate-to-source voltage that a MOSFET with a 5 nm gate oxide can withstand? Assume that the oxide breaks down at 8 MV/cm and the substrate voltage is zero.

SOLUTION

$$V = \mathcal{E} \times d = 8 \times 10^6 \times 5 \times 10^{-7} = 4 \text{ V.}$$

15.3.2 CMOS Technology

Figure 17a shows a CMOS inverter. The gate of the upper PMOS device is connected to the gate of the lower NMOS device. Both devices are enhancement-mode MOSFETs with threshold voltage V_{Tp} less than zero for the PMOS device and V_{Tn} greater than zero for the NMOS device (typically the threshold voltage is about $1/4 V_{DD}$). When the input voltage V_i is zero, the potential V_{GS} of PMOS is $-V_{DD}$ (which is more negative than V_{Tp}), and the NMOS device is off. Hence, the output voltage V_o is very close to V_{DD} (logic 1). When the input is at V_{DD} , the PMOS (with $V_{GS} = 0$) is turned off, and the NMOS is turned on ($V_i = V_{DD} > V_{Tn}$). Therefore, the output voltage V_o equals zero (logic 0). The CMOS inverter has a unique feature: in either logic state, one device in the series path from V_{DD} to ground is nonconductive. The current that flows in either steady state is a small leakage current, and only when both devices are on during switching does a significant current flow through the CMOS inverter. Thus, the average power dissipation is small, on the order of nanowatts. The low power consumption is the most attractive feature of the CMOS circuit. As the number of components per chip increases, power dissipation becomes a major limiting factor.

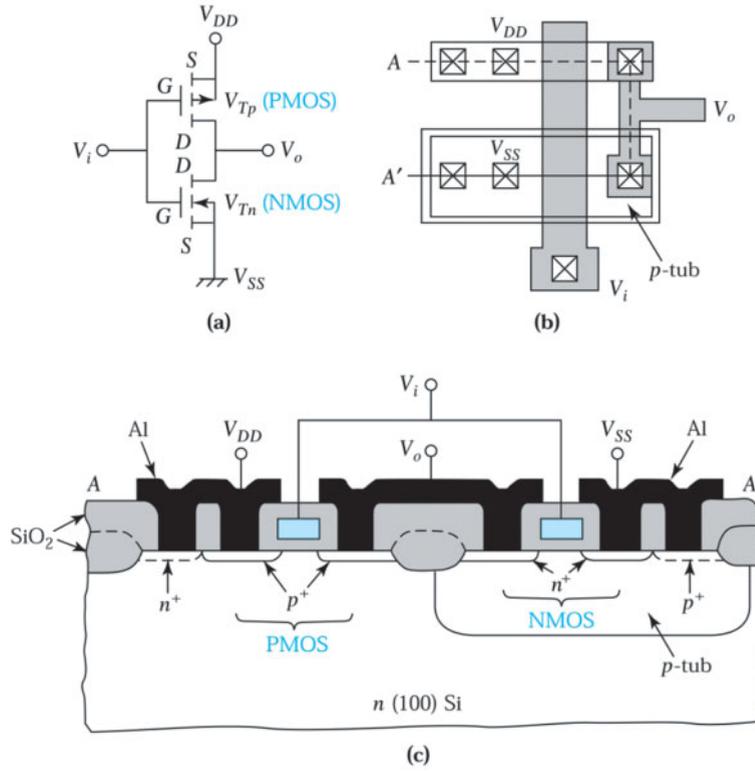


Fig. 17 Complementary MOS (CMOS) inverter. (a) Circuit diagram. (b) Circuit layout. (c) Cross section along dotted $A-A'$ line of (b).

Figure 17b shows a layout of the CMOS inverter and Fig. 17c shows the device cross section along the $A-A'$ line. In processing, a p -tub (also called a p -well) is first implanted and subsequently driven into the n -substrate. The p -type dopant concentration must be high enough to overcompensate the background doping of the n -substrate. The subsequent processes for the n -channel MOSFET in the p -tub are identical to those described previously. For the p -channel MOSFET, $^{11}\text{B}^+$ or $^{49}(\text{BF}_2)^+$ ions are implanted into the n -substrate to form the source and drain regions. A channel implant of $^{75}\text{As}^+$ ions may be used to adjust the threshold voltage and a n^+ -chanstop is formed underneath the field oxide around the p -channel device. Because of the p -tub and the additional steps needed to make the p -channel MOSFET, the number of steps to make a CMOS circuit is essentially double that to make an NMOS circuit. Thus, we have a trade-off between complexity of processing and reduction in power consumption.

Instead of the p -tub described above, an alternate approach is to use an n -tub formed in p -type substrate, as shown in Fig. 18a. In this case, the n -type dopant concentration must be high enough to overcompensate for the background doping of the p -substrate (i.e., $N_D > N_A$). In both the p -tub and the n -tub approach, the channel mobility will be degraded because mobility is determined by the total dopant concentration ($N_A + N_D$). A recent approach using two separated tubs implanted into a lightly doped substrate is shown in Fig. 18b. This structure is called a *twin tub*.¹ Because no overcompensation is needed in either twin tub, higher channel mobility can be attained.

All CMOS circuits have the potential for a troublesome problem called latchup that is associated with parasitic bipolar transistors (for how this problem can occur, see Chapter 6). An effective processing technique to eliminate latchup problem is to use the deep-trench isolation, as shown⁸ in Fig. 18c. In this technique, a trench

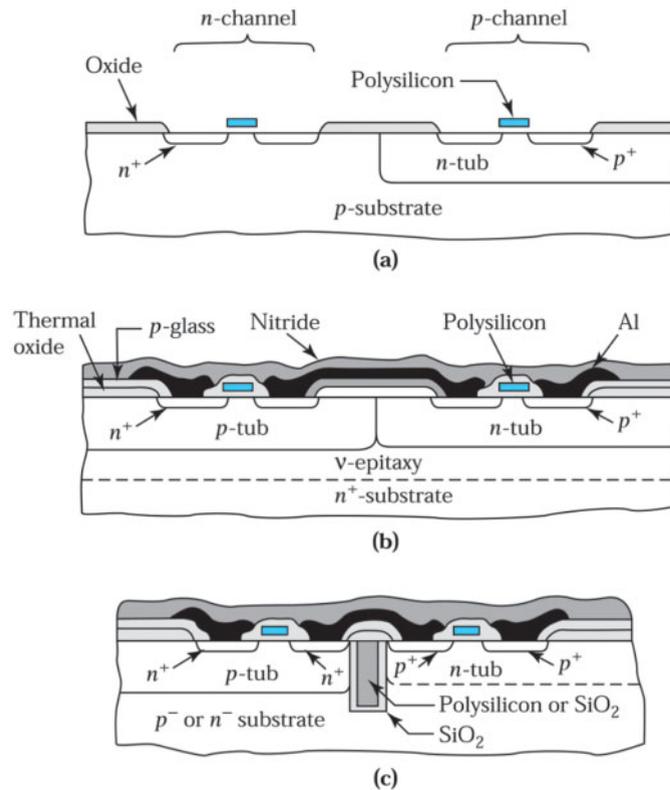


Fig. 18 Various CMOS structures. (a) *n*-tub. (b) Twin tub¹. (c) Refilled trench.⁸

with a depth deeper than the well is formed in the silicon by anisotropic reactive sputter etching. An oxide layer is thermally grown on the bottom and walls of the trench, which is then refilled by deposited polysilicon or silicon dioxide. This technique can eliminate latchup because the *n*-channel and *p*-channel devices are physically isolated by the refilled trench. The detailed steps for some related CMOS processes are now considered.

Well-Formation Technology

The well of a CMOS can be a single well, a twin well, or a retrograde well. The twin-well process has some disadvantages; e.g., it needs high-temperature processing (above 1050°C) and a long diffusion time (more than 8 hours) to achieve the required depth of 2–3 μm. In this process, the doping concentration is highest at the surface and decreases monotonically with depth. To reduce the process temperature and time, high-energy implantation is used, i.e., implanting the ion to the desired depth instead of diffusing from the surface. Since the depth is determined by the implantation energy, we can design the well depth with different implantation energy. The profile of the well in this case can have a peak at a certain depth in the silicon substrate. This is called a retrograde well. Figure 19 compares the impurity profiles in the retrograde well and the conventional thermal diffused well.⁹ The energy for the *n*- and *p*-type retrograde wells is around 700 keV and 400 keV, respectively. As mentioned above, the advantage of the high-energy implantation is that it can form the well under low-temperature and short-time conditions, and hence, can reduce lateral diffusion and increase device density. The retrograde well can offer some additional advantages over the conventional well: (a) because of high doping near the bottom, the well resistivity is lower than that of the conventional well and the latchup problem can be minimized, (b) the chanstop can be formed at the same time as the retrograde well implantation, reducing processing steps and time, and (c) higher well doping in the bottom can reduce the chance of punch through from the drain to the source.

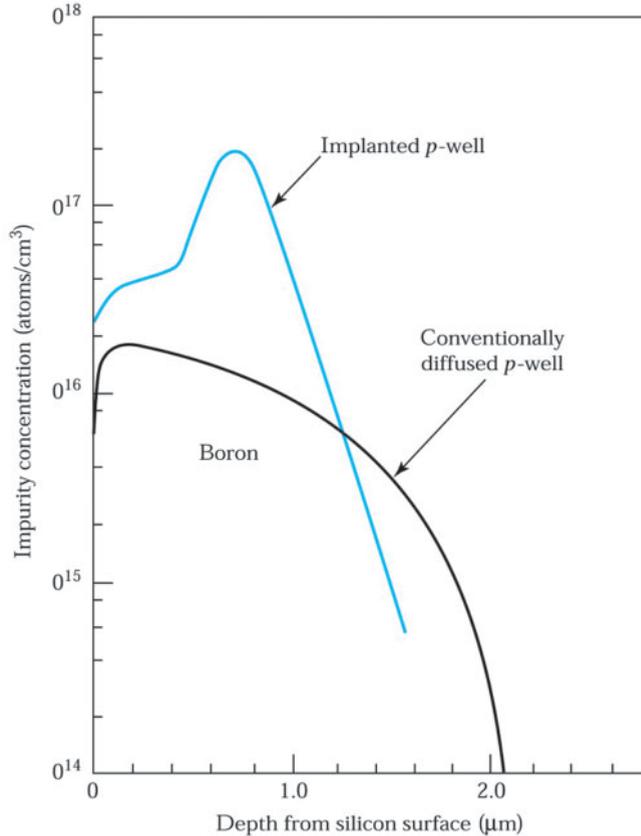


Fig. 19 Retrograded p -well implanted impurity concentration profile. Also shown is a conventionally diffused well.⁹

Gate-Engineering Technology

If we use n^+ -polysilicon for both PMOS and NMOS gates, the threshold voltage for PMOS ($V_{TP} \cong -0.5$ to -1.0 V) has to be adjusted by boron implantation. This makes the channel of the PMOS a buried type, as shown in Fig. 20a. The buried-type PMOS suffers serious short-channel effects as the device size shrinks to $0.25 \mu\text{m}$ and less. The most noticeable phenomena for short-channel effects are the V_T roll-off, drain-induced barrier lowering (DIBL), and the large leakage current at the off state so that even with the gate voltage at zero, leakage current flows through source and drain. To alleviate this problem, one can change n^+ -polysilicon to p^+ -polysilicon for PMOS. Due to the work function difference (there is a 1.0 eV difference from n^+ - to p^+ -polysilicon), one can obtain a surface p -type channel device without the boron V_T adjustment implantation. Hence, as the technology shrinks to $0.25 \mu\text{m}$ and less, dual-gate structures are required, i.e., p^+ -polysilicon gate for PMOS, and n^+ -polysilicon for NMOS (Fig. 20b). A comparison of V_T for the surface channel and buried channel is shown in Fig. 21. We note that the V_T of the surface channel rolls off slowly in the deep-submicron regime compared with the buried-channel device. This makes the surface-channel device with the p^+ -polysilicon suitable for deep-submicron device operation.

To form the p^+ -polysilicon gate, ion implantation of BF_2^+ is commonly used. However, boron penetrates easily from the poly-Si through the oxide into the silicon substrate at high temperatures, resulting in a V_T shift. This penetration is enhanced in the presence of a F-atom. There are methods to reduce this effect: use of rapid thermal annealing to reduce the time at high temperatures and, consequently, the diffusion of boron; use of nitrided oxide to suppress the boron penetration, since boron can easily combine with nitrogen and becomes less mobile; and making a multilayer of polysilicon to trap the boron atoms at the interface of the two layers.

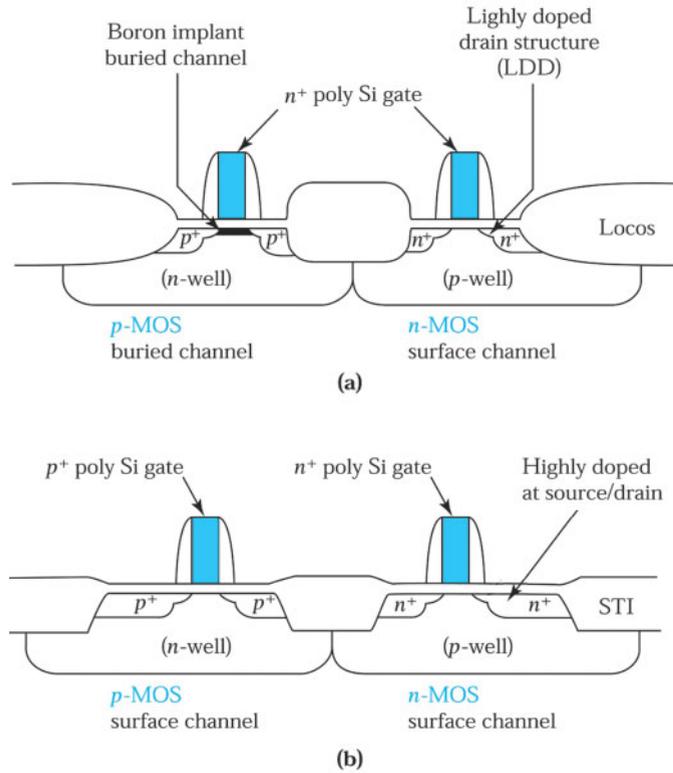


Fig. 20 (a) Conventional long-channel CMOS structure with a single-polysilicon gate (n^+). (b) Advanced CMOS structures with dual-polysilicon gates.

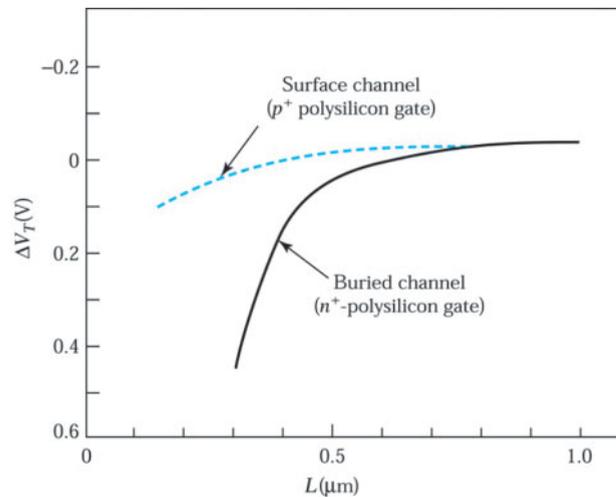


Fig. 21 V_T roll-off for a buried type channel and for a surface type channel. V_T drops very quickly as the channel length becomes less than $0.5 \mu\text{m}$.

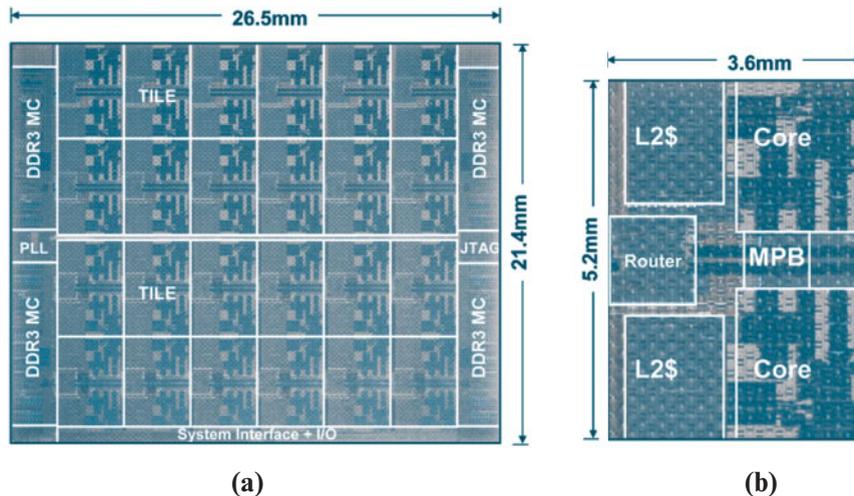


Fig. 22 Micrograph of a 48-core microprocessor, (a) full-chip and (b) tile micrograph. (Photography courtesy of Intel Corporation.)¹⁰

Figure 22 shows a 48-core microprocessor chip that has an area of about 567 mm² and contains 1.3 billion transistors.¹⁰ This ULSI chip is fabricated using 45 nm CMOS technology with a nine-level copper metallization.

15.3.3 BiCMOS Technology

The BiCMOS technology combines both CMOS and bipolar device structures in a single IC. The reason to combine these two different technologies is to create an IC chip with the advantages of both CMOS and bipolar devices. We know that CMOS exhibits advantages in power dissipation, noise margin, and packing density, whereas bipolar has advantages in switching speed, current drive capability, and analog capability. As a result, for a given design rule, BiCMOS can have a greater speed than CMOS, better performance in analog circuits than CMOS, a lower power-dissipation than bipolar, and higher component density than bipolar.

BiCMOS has been widely used in many applications. In the early days, it was used in SRAM. At the present time, BiCMOS technology has been successfully developed for transceiver, amplifier, and oscillator applications in wireless-communication equipment. Most of the BiCMOS processes are based on the CMOS process, with some modifications, such as adding masks for bipolar transistor fabrication. The following example is for a high-performance BiCMOS process based on the twin-well CMOS process, shown¹¹ in Fig. 23.

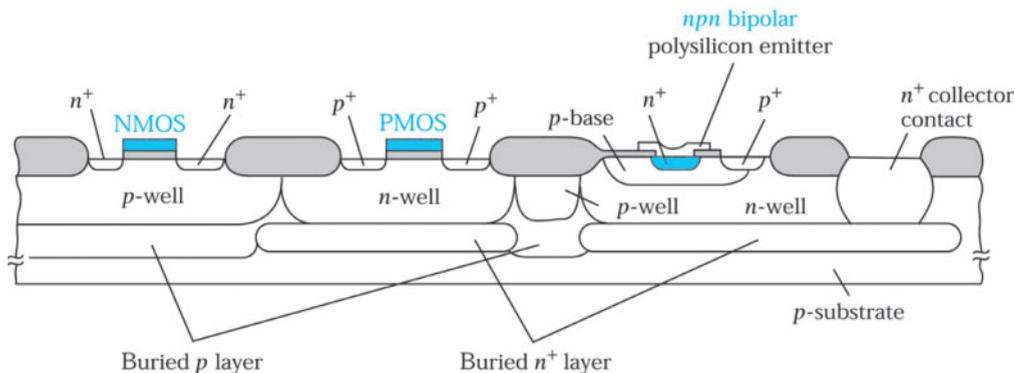


Fig. 23 Optimized BiCMOS device structure. Key features include self-aligned p and n^+ buried layers for improved packing density, separately optimized n - and p -well (twin-well CMOS) formed in an epitaxial layer with intrinsic background doping, and a polysilicon emitter for improved bipolar performance.¹¹

The initial material is a p -type silicon substrate with an n^+ -buried layer formed to reduce the collector's resistance. The buried p -layer is formed through ion implantation to increase the doping level to prevent punchthrough. A lightly doped n -epi layer is grown on the wafer and a twin-well process for the CMOS is performed. To achieve high performance of the bipolar transistor, four additional masks are needed. They are the buried n^+ -mask, the collector deep- n^+ -mask, the base p -mask, and the poly-emitter mask. In other processing steps, the p^+ -region for the base contact can be formed with the p^+ -implant in the source/drain implantation of the PMOS and the n^+ -emitter can be formed with the source/drain implantation of the NMOS. The additional masks and longer processing time than a standard CMOS are the main drawbacks of BiCMOS. The additional cost should be justified, however, by the enhanced performances of BiCMOS.

15.3.4 FinFET Technology

To overcome short channel effects, three-dimensional MOSFETs were developed as discussed in Chapter 6, Section 6.3.3. Among them, FinFET is a typical structure. The device structure of the FinFET is shown in Fig. 24.¹² The channel was formed on the side “vertical” surface of the Si-fin, and the current flows in parallel to the wafer surface. The heart of the FinFET is a thin (~ 10 nm) Si fin that serves as the body of the MOSFET. A heavily doped poly-Si film wraps around the fin and makes electrical contact with its vertical faces. The poly-Si film greatly reduces the source/drain series resistance and provide a convenient means for local interconnect and making connections to the metal. A gap is etched through the poly-Si film to separate the source and drain. The width of this gap, further reduced by the dielectric spacers, determines the gate length. The channel width is basically twice the fin height (plus the fin width). The conducting channel is wrapped around the surface of the fin (hence the name FinFET). Because the source/drain and gate are much thicker (taller) than the fin, the device structure is quasiplanar.

The typical fabrication sequence is shown in Fig. 25.

1. A conventional SOI wafer with a 400-nm thick buried oxide layer and 50-nm thick silicon film can be used as the starting material, except that the alignment notch of the wafer is preferably rotated 45° about the axis of symmetry of the wafer. The reason for this deviation is to provide $\{100\}$ planes on silicon fins.
2. The CVD Si_3N_4 and SiO_2 stack layer is deposited on the silicon film to make a cover layer that will protect the Si-fin through the fabrication process. The fine Si-fin is patterned by electron beam lithography.
3. Phosphorus-doped amorphous Si (for source and drain pads) is deposited at 480°C , and will be polycrystallized in the following step. After amorphous Si deposition, SiO_2 is deposited at 450°C . The process temperatures are low enough to suppress impurity diffusion into the Si fin.

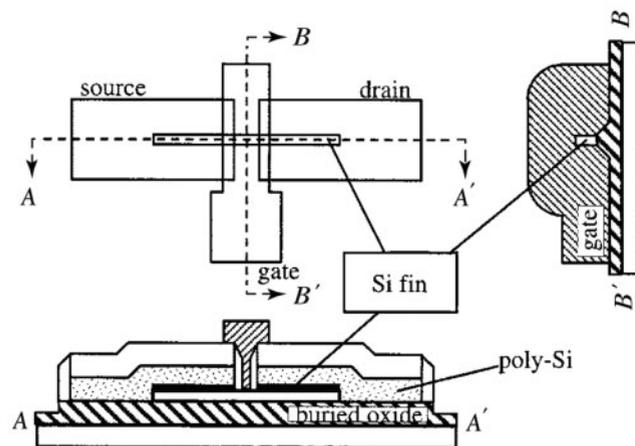


Fig. 24 Schematic parts of a FinFET.¹²

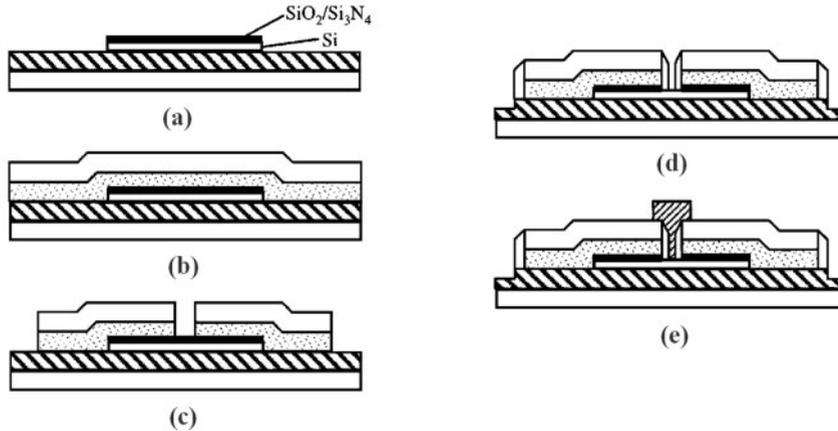


Fig. 25 Process flow of FinFET: (a) after depositing the Si_3N_4 and SiO_2 stacked layer, the Si fin is formed; (b) a phosphorus-doped-poly Si and SiO_2 stacked layer is deposited; (c) the source and drain are etched while the Si fin is covered by the mask layer; (d) a SiO_2 space layer is etched into the buried oxide layer; and (e) after depositing B-doped SiGe, the gate pattern is delineated.¹²

4. Using electron-beam lithography, the S/D pads with a narrow gap in between them are delineated. The SiO_2 and amorphous Si layers are etched and the gap is formed. While the cover layer protects the Si fin, the amorphous Si is completely removed from the side of the Si fin. The amorphous Si in contact with the Si fin at its side surfaces becomes the impurity diffusion source that forms the transistor S/D later.
5. CVD SiO_2 is deposited to make spacers around the S/D pads. The height of the Si fin is 50 nm, and the total pads thickness is 400 nm. Making use of the difference in the heights, the SiO_2 spacer on the sides of the Si-fin is completely removed by over etching of SiO_2 while the cover layer protects the Si-fin. The Si surface is again exposed on the sides of the Si fin. During this over-etching, SiO_2 on the S/D pads and the buried oxide between S/D pads are etched.
6. By oxidizing the Si surface, gate oxide as thin as 2.5 nm is grown. During gate oxidation, the amorphous Si of the S/D pads is crystallized. Also, phosphorus diffuses from the S/D pads into the Si fin and forms the S/D extensions under the oxide spacers. Then, the gate deposition follows.

15.3.5 Memory Devices

Memories are devices that can store digital information (or data) in terms of *bits* (binary digits). Various memory chips have been designed and fabricated using CMOS technology. MOS memory structures were introduced in Chapter 6, Section 6.4. In a RAM, memory cells are organized in a matrix structure and data can be accessed (i.e., stored, retrieved, or erased) in random order, independent of their physical locations. A static random access memory (SRAM) can retain stored data indefinitely as long as the power supply is on. The SRAM is basically a flip-flop circuit that can store one bit of information. A SRAM cell has four NMOSFETs and two PMOSFETs in CMOS technology.¹³

To reduce the cell area and power consumption, the dynamic random access memory (DRAM) has been developed. Figure 26a shows the circuit diagram of the one-transistor DRAM cell in which the transistor serves as a switch and one bit of information can be stored in the storage capacitor. The voltage level on the capacitor determines the state of the cell. For example, +1.5 V may be defined as logic 1 and 0 V defined as logic 0. The stored charge will typically be removed in a few milliseconds mainly because of the leakage current of the capacitors; thus dynamic memories require periodic “refreshing” of the stored charge.

Figure 26b shows the layout of a DRAM cell and Fig. 26c shows the corresponding cross section through AA'. The storage capacitor uses the channel region as one plate, the polysilicon gate as the other plate, and

the gate oxide as the dielectric. The row line is a metal track to minimize the delay due to parasitic resistance (R) and parasitic capacitance (C), the RC delay. The column line is formed by n^+ -diffusion. The internal drain region of the MOSFET serves as a conductive link between the inversion layers under the storage gate and the transfer gate. The drain region can be eliminated by using the double-level polysilicon approach shown in Fig. 26*d*. The second polysilicon electrode is separated from the first polysilicon capacitor plate by an oxide layer that is thermally grown on the first-level polysilicon before the second electrode has been defined. The charge from the column line can therefore be transmitted directly to the area under the storage gate by the continuity of inversion layers under the transfer and storage gates.

To meet the requirements of high-density DRAM, the DRAM structure has been extended to the third dimension with stacked or trench capacitors. Figure 27*a* shows a simple trench cell structure.¹⁴ The advantage of the trench type is that the capacitance of the cell can be increased by increasing the depth of the trench without increasing the surface area of silicon occupied by the cell. The main difficulties of making trench-type cells are the etching of the deep trench, which needs a rounded bottom corner and growth of a uniform, thin dielectric film on the trench walls. Figure 27*b* shows a stacked cell structure. The storage capacitance increases as a result of stacking the storage capacitor on top of the access transistor. The dielectric is formed using the thermal oxidation or CVD nitride methods between the two-polysilicon plates. Hence, the stacked cell process is easier than the trench type process.

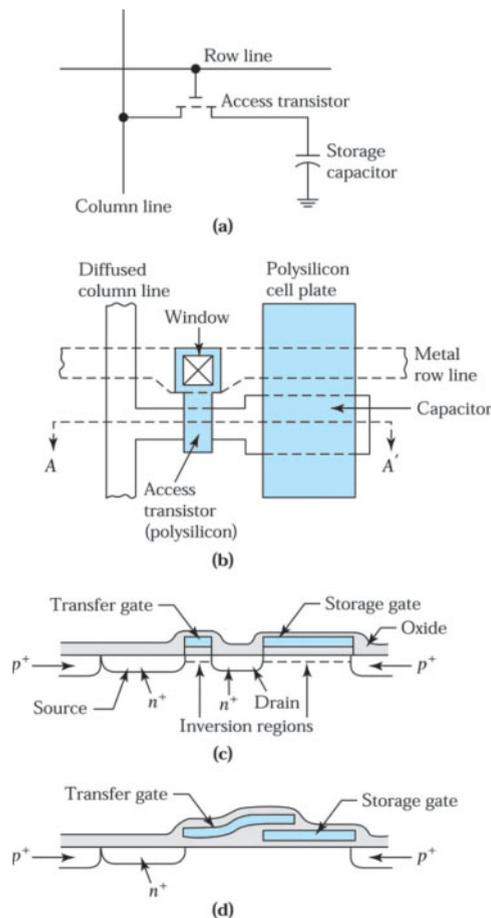


Fig. 26 Single-transistor dynamic random access memory (DRAM) cell with a storage capacitor.¹³ (a) Circuit diagram. (b) Cell layout. (c) Cross section through A-A'. (d) Double-level polysilicon.

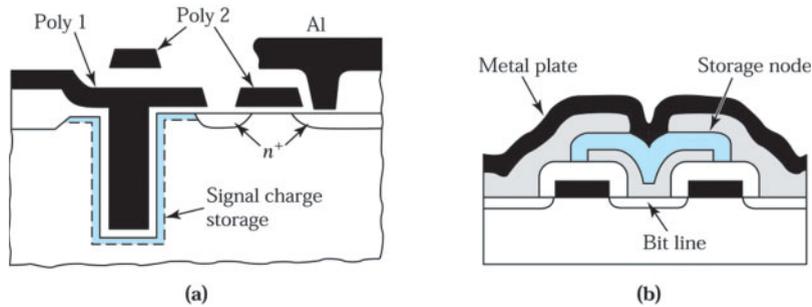


Fig. 27 (a) DRAM with a trench cell structure.¹⁴ (b) DRAM with a single-layer stacked-capacitor cell.

Figure 28 shows an 8 Gb DRAM chip.¹⁵ This memory chip uses 50 nm fabrication process for the high-speed, low-power DRAM. The memory chip has an area of 98 mm² and operates at 1.5 V. Low-resistance copper wiring and low-dielectric-constant film ($k = 2.96$) are used for the wiring.

Both SRAM and DRAM are volatile memories: that is, they lose their stored data when power is switched off. Nonvolatile memories discussed in detail in Section 6.4.3 of Chapter 6, on the other hand, can retain their data. A floating-gate nonvolatile memory is basically a conventional MOSFET that has a modified gate electrode. The composite gate has a regular (control) gate and a floating gate that is surrounded by insulators. When a large positive voltage is applied to the control gate, charge will be injected from the channel region through the gate oxide into the floating gate. When the applied voltage is removed, the injected charge can be stored in the floating gate for a long time. To remove this charge, a large negative voltage must be applied to the control gate, so that the charge will be injected back into the channel region.

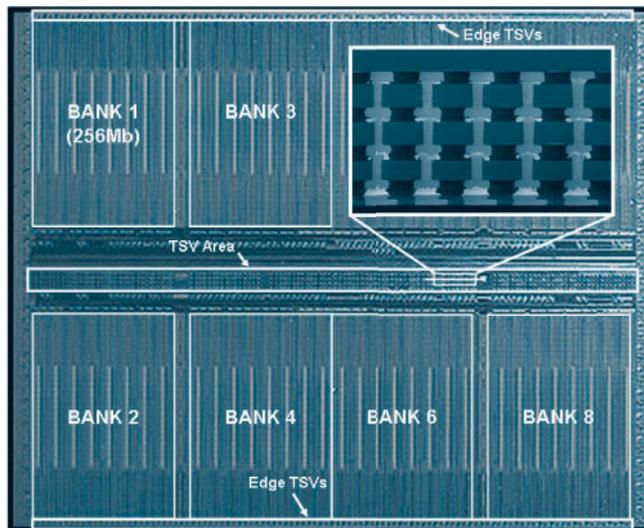


Fig. 28 An 8 Gb DRAM that contains over 16 billion components. (Photography courtesy of Samsung Electronics.)¹⁵

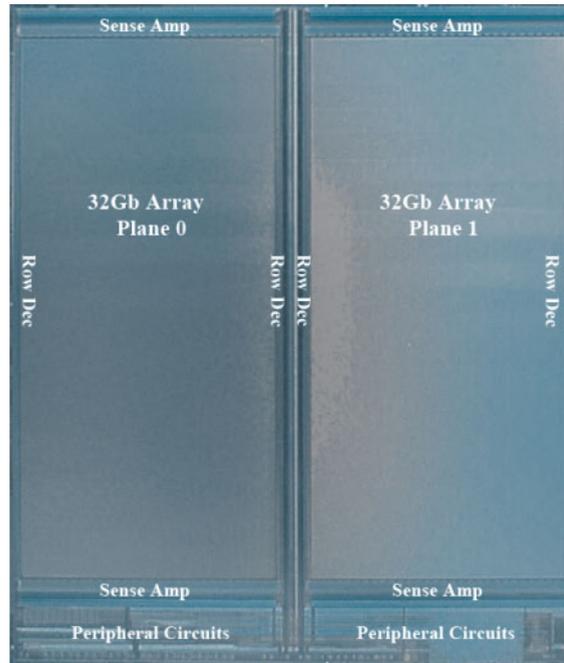


Fig. 29 A 5.6 MB/s, 64 Gb, 4b/cell NAND flash memory. (Photography courtesy of SanDisk/Toshiba.)¹⁶

Another version of nonvolatile memory is the metal-nitride-oxide-semiconductor (MNOS) type also discussed in Section 6.4.3 of Chapter 6. When a positive gate voltage is applied, electrons can tunnel through the thin oxide layer (~2 nm) and be captured by the traps at the oxide-nitride interface, thus becoming stored charges there. The charge stored in the capacitor C causes a shift in the threshold voltage, and the device remains at the higher threshold-voltage state (logic 1). In a well designed memory device, the charge retention time can be over 100 years. To erase the memory (the stored charge) and return the device to a lower threshold-voltage state (logic 0), a gate voltage or other means (such as ultraviolet light) can be used.

The nonvolatile semiconductor memory (NVSM) has been extensively used in portable electronic systems, such as cellular phones and the digital cameras. Another interesting application is the chip card, also called an IC card. Figure 29 shows a 5.6 MB/s, 64 Gb, 4b/cell NAND flash memory.¹⁶ In contrast to the limited volume (1 kbytes) inside a conventional magnetic tape card, the size of the nonvolatile memory can be increased depending on the application (e.g., you can store personal photos or fingerprints). Through IC card read/write machines, the data can be used in numerous applications, such as telecommunications (card telephone, mobile radio), payment transactions (electronic purse, credit card), pay television, transport (electronic ticket, public transport), health care (patient-data card), and access control. The IC card will play a central role in the global information and service society of the future.¹⁷

► 15.4 MESFET TECHNOLOGY

Recent advances in gallium arsenide processing techniques in conjunction with new fabrication and circuit approaches have made possible the development of “silicon-like” gallium arsenide IC technology. There are three inherent advantages of gallium arsenide compared with silicon: higher electron mobility, which results in lower series resistance for a given device geometry; higher drift velocity at a given electric field, which improves device speed; and the ability to be made semi-insulating, which can provide a lattice-matched dielectric-insulated substrate. However, gallium arsenide also has three disadvantages: a very short minority-carrier lifetime; lack of a stable, passivating native oxide; and crystal defects that are many orders of magnitude greater than in silicon.

The short minority-carrier lifetime and the lack of high-quality insulating films have prevented the development of bipolar devices and delayed MOS technology using gallium arsenide. Thus, the emphasis of gallium arsenide IC technology is in the MESFET area, in which our main concerns are majority-carrier transport and the metal-semiconductor contact.

High-performance MESFET structures fall into two major categories: the recessed-channel (or recessed-gate) structure and the ion-implanted planar structure. A typical fabrication sequence^{18,19} for a recessed-channel MESFET is shown in Fig. 30. The active layers are grown epitaxially over the semi-insulating substrate (Fig. 30a). An intrinsic buffer layer is first grown and followed by an active n -channel layer. The buffer layer serves to eliminate defects from the semi-insulating substrate. Finally, an epitaxial n^+ -contact is grown over the active

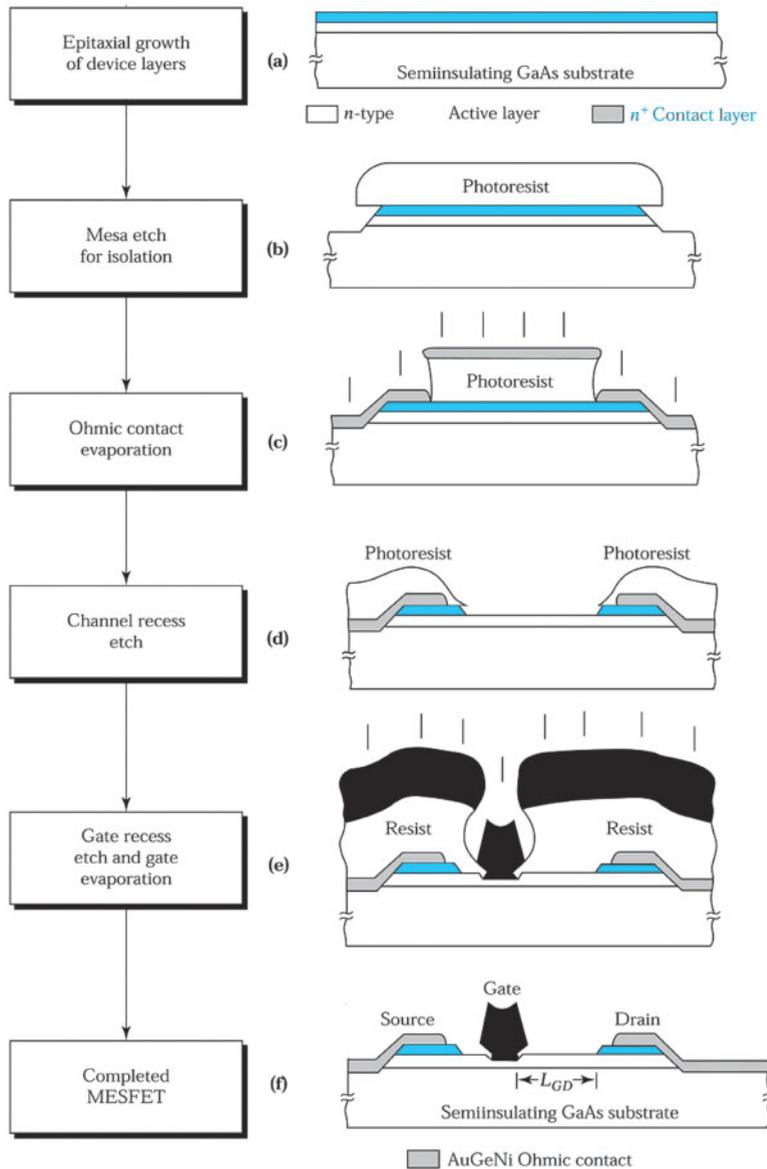


Fig. 30 Fabrication sequence of a GaAs MESFET.¹⁸

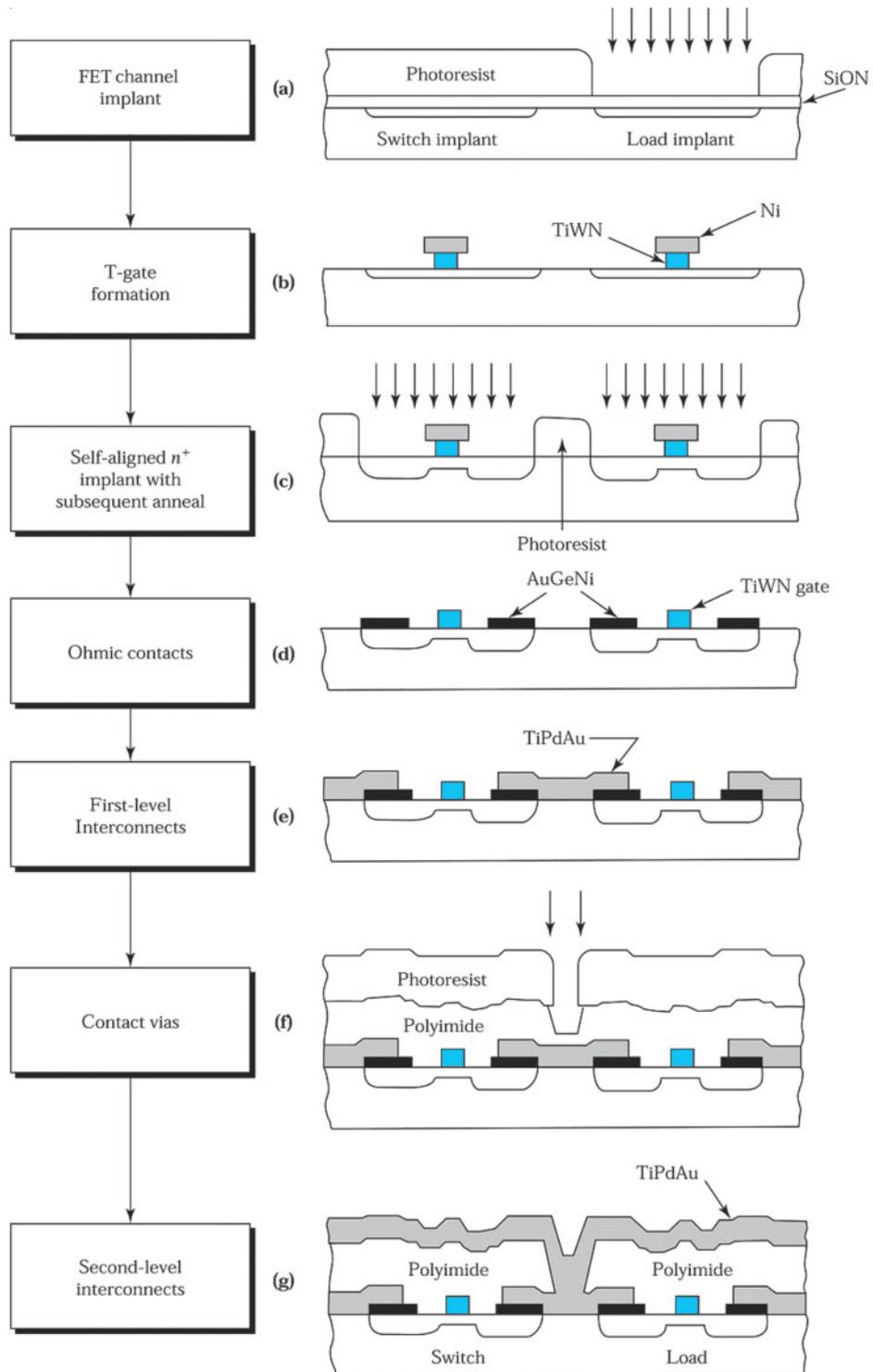


Fig. 31 Fabrication process for MESFET direct-coupled FET logic (DCFL) with active loads. Note that the n^+ -source and drain regions are self-aligned to the gate.²⁰

n -channel to reduce the source and drain contact resistance. A mesa-etch step is performed for isolation (Fig. 30*b*), and a metal layer is evaporated for the source and drain ohmic contacts (Fig. 30*c*). A channel-recess etch is followed by a gate-recess etch and gate evaporation (Fig. 30*d* and *e*). Sometimes this etching process is monitored by measuring the current between source and drain for more precise control of the final channel current. One of the advantages of this recessed-channel structure is that the surface is further away from the n -channel layer so that surface effects such as transient response and other reliability problems are minimized. One disadvantage of this scheme is the additional steps required for isolation, which could be a mesa-etching process (as shown) or an isolation implantation that converts the semiconductor into high-resistivity material. After a liftoff process that removes the photoresist, as shown in Fig. 30*e*, the MESFET is completed (Fig. 30*f*).

Note that the gate is offset toward the source to minimize the source resistance. The epitaxial layer is thick enough to minimize the effect of surface depletion on the source and drain resistance. The gate has the shape of a T-gate or mushroom-gate. The shorter dimension at the bottom of the gate is the electrical channel length and serves to optimize f_T and g_m , while the wider top portion reduces the gate resistance for an improved f_{max} . In addition, the length L_{GD} is designed to be greater than the depletion width at gate-drain breakdown.

The ion-implanted planar structure is useful for the fabrication of MESFET integrated circuit as shown²⁰ in Fig. 31. The active region is created by ion implantation to overcompensate the deep-level impurities in the semi-insulating substrate. A relatively light channel implant is used for the enhancement-mode switching device and a heavier implant is used for the depletion-mode load device. To minimize the source and drain parasitic resistance, the deeper n^+ -implantation source and drain regions should be as close to the gate as possible. This is done by various self-aligned processes. In a gate-priority self-aligned process, the gate is formed first and the source/drain ion implantation is self-aligned to the gate. In this process, since ion implantation requires high-temperature annealing for activation, the gate must be made of materials that can withstand high-temperature processing. Examples are Ti-W alloy (for example TiWN), WSi_2 , and $TaSi_2$. The second approach is ohmic-priority, where the source/drain implantation and anneal are done before the gate formation. This process relaxes the previous requirement on the gate material. A gate recess is usually not used for such digital IC fabrication because the uniformity of each depth is difficult to control, leading to unacceptable variation in the threshold voltage. This process sequence can also be used for a monolithic microwave integrated circuit (MMIC). Note that the gallium arsenide MESFET processing technology is similar to the silicon-based MOSFET processing technology.

Gallium arsenide ICs with complexities up to the large-scale integration level ($\sim 10,000$ components per chip) have been fabricated. Because of the higher drift velocity ($\sim 20\%$ higher than silicon), gallium arsenide ICs will have a 20% greater speed than silicon ICs that use the same design rules. However, substantial improvements in crystal quality and processing technology are needed before gallium arsenide can seriously challenge the preeminent position of silicon in ULSI applications.

► 15.5 CHALLENGES FOR NANOELECTRONICS

Since the beginning of the integrated-circuit era in 1959, the minimum device dimension, also called the minimum feature length, has been reduced at an annual rate of about 13% (i.e., a reduction of 35% every three years). According to a prediction of the International Technology Roadmap for Semiconductors,²¹ the minimum feature length will shrink from 130 nm in 2002 to 35 nm around 2014, as shown in Table 1. Also shown in Table 1 is the DRAM size. The DRAM quadruples its memory cell capacity every three years, and 64 Gbit DRAM will be available in 2011 using 50 nm design rules. The table also shows that wafer size will increase to 450 mm (18 in. diameter) in 2014. In addition to the feature size reduction, challenges come at the device level, material level, and system level, as discussed in the following sections.

15.5.1 Challenges for Integration

Figure 32 shows the trends in power supply voltage V_{DD} , threshold voltage V_T , and gate oxide thickness d versus channel length for CMOS logic technology.²² From the figure, one can find that the gate oxide thickness will soon approach the tunneling-current limit of 2 nm. V_{DD} scaling will slow down because of non-scalable V_T (i.e., to

TABLE 1 The Technology Generation²¹ from 1997 to 2014

| Year of first product shipment | 1997 | 1999 | 2002 | 2005 | 2008 | 2011 | 2014 |
|--------------------------------|--------|---------|---------|---------|-------|------|------|
| Feature size (nm) | 250 | 180 | 130 | 100 | 70 | 50 | 35 |
| DRAM size (bit) | 256M | 1G | — | 8G | — | 64G | — |
| Wafer size (mm) | 200 | 300 | 300 | 300 | 300 | 300 | 450 |
| Gate oxide (nm) | 3–4 | 1.9–2.5 | 1.3–1.7 | 0.9–1.1 | <1.0 | — | — |
| Junction depth (nm) | 50–100 | 42–70 | 25–43 | 20–33 | 15–30 | — | — |

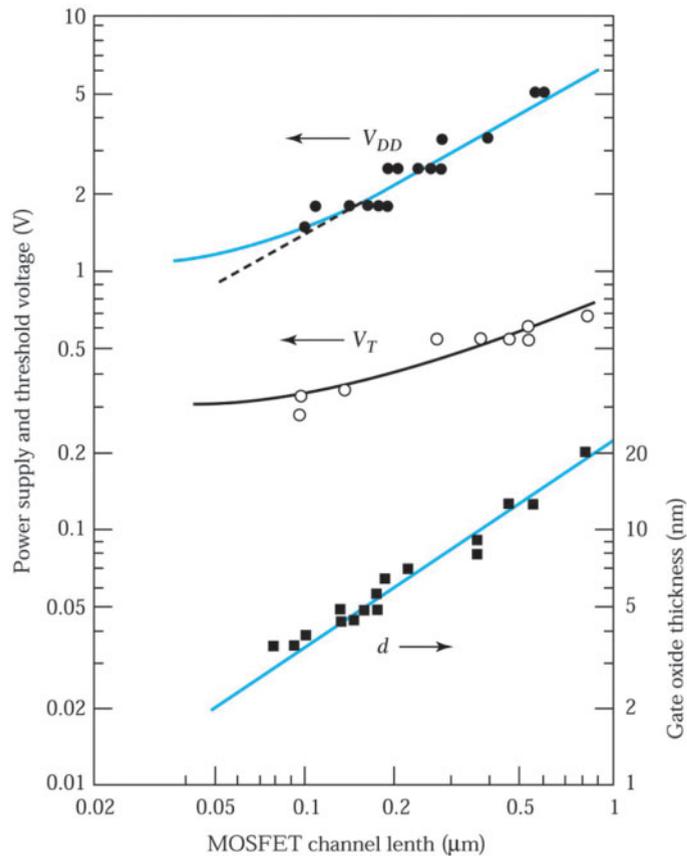


Fig. 32 Trends of power supply voltage V_{DD} , threshold voltage V_T , and gate oxide thickness d versus channel length for CMOS logic technologies. Points are collected from data published over recent years.²²

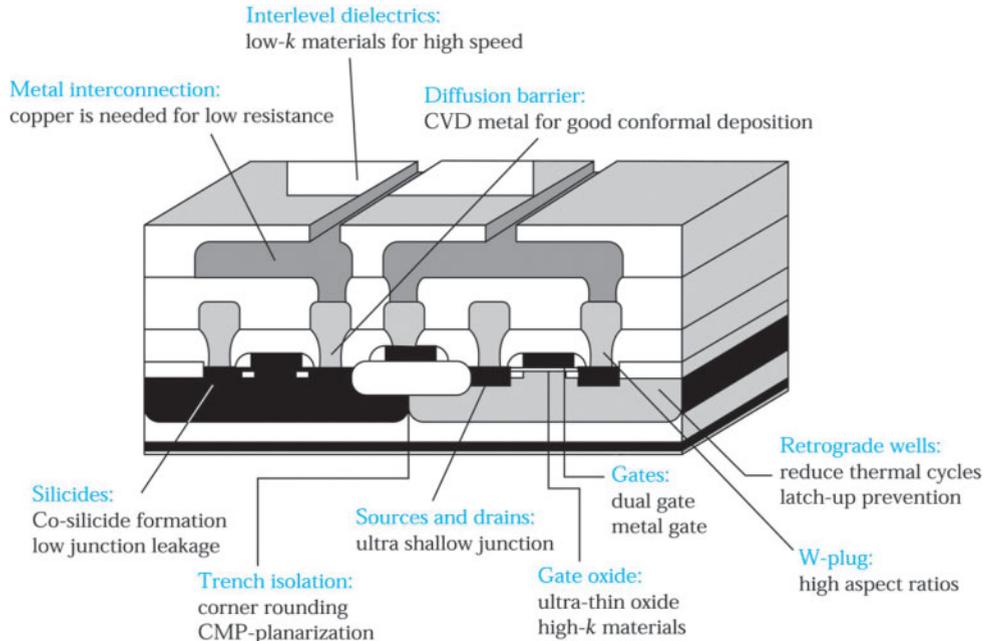


Fig. 33 Challenges for 180 nm and smaller MOSFET.²³

a minimum V_T of about 0.3 V due to subthreshold leakage and circuit noise immunity). Some challenges of the 180 nm technology and beyond are shown²³ in Fig. 33. The most stringent requirements are as follows.

Ultrashallow Junction Formation

As mentioned in Chapter 6, the short-channel effect happens as the channel length is reduced. This problem becomes critical as the device dimension is scaled down to 100 nm. To achieve an ultrashallow junction with low sheet resistance, low-energy (less than 1 keV) implantation technology with high dosage must be employed to reduce the short-channel effect. Table 1 shows the required junction depth versus technology generation. The requirements of junction depths for 100 nm are around 20–33 nm with a doping concentration of $1 \times 10^{20}/\text{cm}^3$.

Ultrathin Oxide

As the gate length shrinks below 100 nm, the oxide equivalent thickness of gate dielectric must be reduced to around 2 nm to maintain performance. However, if only SiO_2 (with a dielectric constant of 3.9) is used, leakage through the gate becomes very high because of direct tunneling. For this reason, thicker high- k dielectric materials that have lower leakage current are used to replace oxide. Candidates for the short term are silicon nitride (with a dielectric constant of 7), Ta_2O_5 (25), HfO_2 (20–25), and TiO_2 (60–100).

Silicide Formation

Silicide-related technology has become an integral part of submicron devices for reducing parasitic resistance to improve device and circuit performance. The conventional Ti-silicide process has been widely used in 350–250 nm technology. However, the sheet resistance of a TiSi_2 line increases with decreasing line width, which limits the use of TiSi_2 in 180 nm CMOS applications and beyond. CoSi_2 or NiSi processes will replace TiSi_2 in the technology beyond 180 nm.

New Materials for Interconnection

To achieve high-speed operation, the RC time delay of the interconnection must be reduced.²⁴ Figure 12 of Chapter 12 showed the delay as a function of feature size. It is obvious that the gate delay decreases as the channel length decreases; meanwhile, the delay resulting from interconnect increases significantly as the size decreases. This causes the total delay time to increase as the dimension of the device size scales below 100 nm. Consequently, both high-conductivity metals, such as Cu, and low-dielectric-constant (low- k) insulators, such as organic (polyimide) or inorganic (F-doped oxide) materials offer major performance gains. Cu exhibits superior performance because of its high conductivity ($1.7 \mu\Omega\text{-cm}$ compared with $2.7 \mu\Omega\text{-cm}$ of Al) and is 10–100 times more resistant to electromigration. The delay using the Cu and low- k material shows a significant decrease compared with that of the conventional Al and oxide. Hence, Cu with the low- k material is essential in multilevel interconnection for future deep-submicron technology.

Power Limitations

The power required merely to charge and discharge circuit nodes in an IC is proportional to the number of gates and the frequency at which they are switched (clock frequency). The power can be expressed as $P \cong 1/2 CV^2 nf$, where C is the capacitance per device, V is the applied voltage, n is the number of devices per chip, and f is the clock frequency. The temperature rise caused by this power dissipation in an IC package is limited by the thermal conductivity of the package material, unless auxiliary liquid or gas cooling is used. The maximum allowable temperature rise is limited by the bandgap of the semiconductor ($\sim 100^\circ\text{C}$ for Si with a bandgap of 1.1 eV). For such a temperature rise, the maximum power dissipation of a typical high-performance package is about 10 W. As a result, we must limit either the maximum clock rate or the number of gates on a chip. As an example, in an IC containing 100 nm MOS devices with $C = 5 \times 10^{-2}$ fF, running at a 20 GHz clock rate, the maximum number of gates we can have is about 10^7 if we assume a 10% duty cycle. This is a design constraint fixed by basic material parameters.

SOI Integration

The isolation of the SOI was mentioned in Chapter 15, Section 15.2.2. Recently SOI technology has received more attention. The advantages of SOI integration become significant as the minimum feature length is reduced below 100 nm. From the process point of view, SOI does not need the complex well structure and isolation processes. In addition, shallow junctions are directly obtained through the SOI film thickness. There is no risk of nonuniform interdiffusion of silicon and Al in the contact regions because of oxide isolation at the bottom of the junction. Hence, the contact barrier is not necessary. From the device point of view, the modern bulk silicon device needs high doping at the drain and substrate to eliminate short-channel effects and punch-through. This high doping results in high capacitance when the junction is reversed bias. On the other hand, in SOI, the maximum capacitance between the junction and substrate is the capacitance of the buried insulator whose dielectric constant is three times smaller than that of silicon (3.9 versus 11.9). Based on the ring oscillator performance, the 130 nm SOI CMOS technology can achieve 25% faster speed or require 50% less power than a similar bulk technology.²⁵ SRAM, DRAM, CPU, and rf CMOS have all been successfully fabricated using SOI technology. Therefore SOI is a key candidate for the future system-on-a-chip technology, considered in the following section.

► EXAMPLE 5

For an equivalent oxide thickness of 1.5 nm, what will be the physical thickness when high- k materials nitride ($\epsilon_{ox}/\epsilon_0 = 7$), Ta_2O_5 (25), or TiO_2 (80) is used?

SOLUTION For nitride,

$$\begin{aligned} \left(\frac{\epsilon_{ox}}{1.5} \right) &= \left(\frac{\epsilon_{\text{nitride}}}{d_{\text{nitride}}} \right), \\ d_{\text{nitride}} &= 1.5 \left(\frac{7}{3.9} \right) = 2.69 \text{ nm}. \end{aligned}$$

Using the same calculation, we obtain 9.62 nm for Ta_2O_5 and 10.77 nm for TiO_2 . ◀

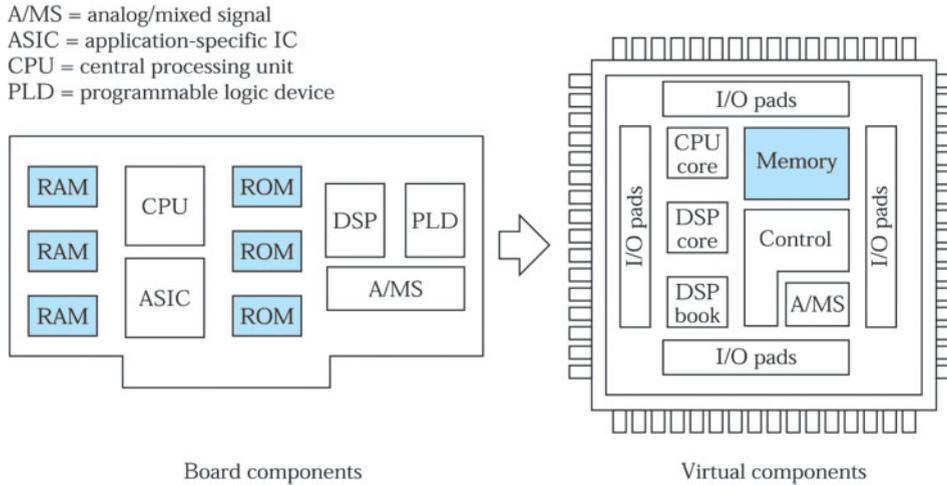


Fig. 34 System-on-a-chip of a conventional personal-computer motherboard.²⁶

15.5.2 System-on-a-Chip

The increased component density and improved fabrication technology have helped the realization of the system-on-a-chip (SOC), that is, an IC chip that contains a complete electronic system. The designers can build all the circuitry needed for a complete electronic system, such as a camera, radio, television, or personal computer (PC), on a single chip. Figure 34 shows the SOC application in the PC's motherboard. Components (11 chips in this case) once found on boards have become virtual components on the chip at the right.²⁶ In addition, the system-on-a-chip can be integrated into 3-D system integration²⁷ and can have higher-level functions.

There are two obstacles to the realization of the SOC. The first is the huge complexity of the design. Since the component board is presently designed by different companies and different design tools, it is difficult to integrate them into one chip. The other is the difficulty of fabrication. In general, fabricating processes for the DRAM are significantly different from those of logic IC (e.g., CPU). Speed is the first priority for the logic, whereas leakage of the stored charge is the priority for memory. Therefore, multilevel interconnection using five to eight levels of metals is essential for logic IC to improve the speed. However, DRAM needs only two to three levels. In addition, to increase the speed, a silicide process must be used to reduce the series resistance, and ultrathin gate oxide is needed to increase the drive current. These requirements are not critical for the memory.

To achieve the SOC goal, an embedded DRAM technology is introduced, i.e., to merge logic and DRAM into a single chip with compatible processes. Figure 35 shows the schematic cross section of the embedded DRAM, including the DRAM cells and the logic CMOS devices.²⁸ Some processing steps are modified as a compromise. The trench-type capacitor is used instead of the stacked type so that there is no height difference in the DRAM cell structure. In addition, multiple gate oxide thicknesses exist on the same wafer to accommodate multiple supply voltages and/or combine memory and logic circuits on one chip.

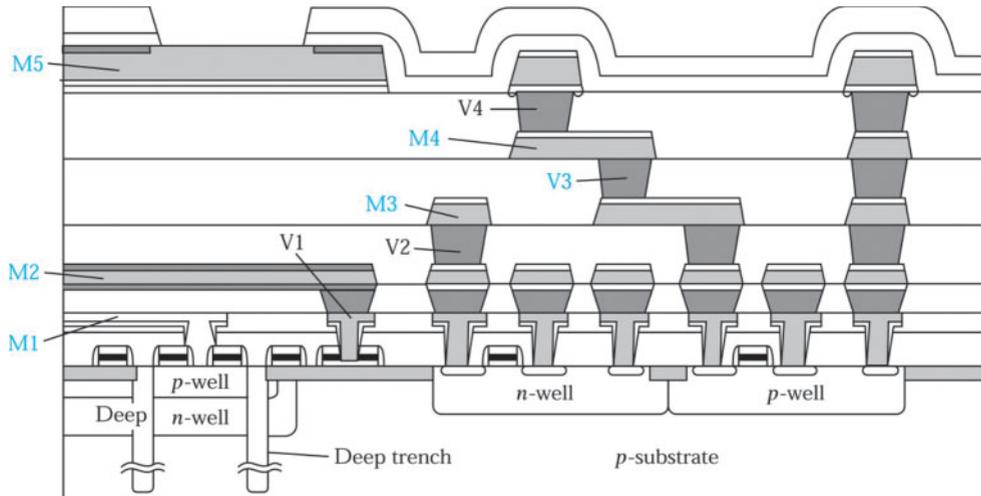


Fig. 35 Schematic cross section of the embedded DRAM including DRAM cells and logic MOSFETs. There is no height difference in the trench capacitor cell because of the DRAM cell structure. $M1$ to $M5$ are metal interconnections and $V1$ to $V4$ are via holes.²⁸

► SUMMARY

In this chapter we considered processing technologies for passive components, active devices, and ICs. Three major IC technologies based on the bipolar transistor, the MOSFET, and the MESFET were discussed in detail. It appears that the MOSFET will be the dominant technology in the foreseeable future because of its superior performance compared with the bipolar transistor. For sub-100 nm CMOS technology, a good candidate is the combination of an SOI-substrate with interconnections using Cu and low- k materials.

Because the rapid reduction in feature length, the technology will soon reach its practical limit as the channel length is reduced to about 20 nm. What the device beyond the CMOS will be is the question being asked by research scientists. Major candidates include many innovative devices based on quantum mechanical effects. This is because when the lateral dimension is reduced to below 100 nm, depending on the materials and the temperature of operation, electronic structures will exhibit nonclassical behaviors. The operation of such devices will be on the scale of single-electron transport. This approach has been demonstrated by the single-electron memory cell. The realization of such systems with trillions of components will be a major challenge beyond CMOS.²⁹

► REFERENCES

1. For a detailed discussion on IC process integration, see C. Y. Liu and W. Y. Lee, "Process Integration," in C. Y. Chang and S. M. Sze, Eds., *ULSI Technology*, McGraw-Hill, New York, 1996.
2. T. Tachikawa, "Assembly and Packaging," in C. Y. Chang and S. M. Sze, Eds., *ULSI Technology*, McGraw-Hill, New York, 1996.

3. T. H. Lee, *The Design of CMOS Radio-Frequency Integrated Circuits*, Cambridge Univ. Press, Cambridge, U.K., 1998, Ch. 2.
4. D. Rise, "Isoplanar-S Scales Down for New Heights in Performance," *Electronics*, **53**, 137 (1979).
5. T. C. Chen et al., "A Submicrometer High-Performance Bipolar Technology," *IEEE Electron. Device Lett.*, **10** (8), 364, (1989).
6. G. P. Li et al., "An Advanced High-Performance Trench-Isolated Self-Aligned Bipolar Technology," *IEEE Trans. Electron Devices*, **34**(10), 2246 (1987).
7. W. E. Beasle, J. C. C. Tsai, and R. D. Plummer, Eds., *Quick Reference Manual for Semiconductor Engineering*, Wiley, New York, 1985.
8. R. D. Rung, H. Momose, and Y. Nagakubo, "Deep Trench Isolation CMOS Devices," *IEEE Tech. Dig. Int. Electron. Devices Meet.*, p. 237, 1982.
9. D. M. Bron, M. Ghezzi, and J. M. Primbley, "Trends in Advanced CMOS Process Technology," *Proc. IEEE*, p. 1646, (1986).
10. J. Howard et al., "A 48-Core IA-32 Message-Passing Processor with DVFS in 45nm CMOS", *Int. Solid-State Circuits Conference*, p.108, 2010.
11. H. Higuchi et al., "Performance and Structure of Scaled-Down Bipolar Devices Merge with CMOSFETs," *IEEE Tech. Dig. Int. Electron. Devices Meet.*, 694, 1984.
12. D. Hisamoto, W. C. Lee, J. Kedzierski, H. Takeuchi, K. Asano, C. Kuo, E. Anderson, T. J. King, J. Bokor, and C. Hu, "FinFET—A Self-Aligned Double-Gate MOSFET Scalable to 20 nm," *IEEE Trans. Electron. Devices*, **47**, 2320 (2000).
13. R. W. Hunt, "Memory Design and Technology," in M. J. Howes and D. V. Morgan, Eds., *Large Scale Integration*, Wiley, New York, 1981.
14. A. K. Sharma, *Semiconductor Memories—Technology, Testing, and Reliability*, IEEE, New York, 1997.
15. U. Kang et al., "8Gb 3D DDR3 DRAM Using Through-Silicon-Via Technology", *Int. Solid-State Circuits Conference*, p.130, 2009.
16. C. Trinh et al., "A 5.6 MB/s 64Gb 4b/Cell NAND Flash Memory in 43nm CMOS", *Int. Solid-State Circuits Conference*, p.246, 2009.
17. U. Hamann, "Chip Cards—The Application Revolution," *IEEE Tech. Dig. Int. Electron. Devices Meet.*, p. 15, 1997.
18. M. A. Hollis and R. A. Murphy, "Homogeneous Field-Effect Transistors," in S. M. Sze, Ed., *High-Speed Semiconductor Devices*, Wiley, New York, 1990.
19. S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, 3rd Ed., Wiley Interscience, Hoboken, 2007.
20. H. P. Singh et al., "GaAs Low Power Integrated Circuits for a High Speed Digital Signal Processor," *IEEE Trans. Electron Devices*, **36**, 240 (1989).
21. *International Technology Roadmap for Semiconductor (ITRS)*, Semiconductor Ind. Assoc., San Jose, 1999.
22. Y. Taur and E. J. Nowak, "CMOS Devices Below 0.1 μm : How High Will Performance Go?" *IEEE Tech. Dig. Int. Electron. Devices Meet.*, 215, 1997.

23. L. Peters, "Is the 0.18 μm Node Just a Roadside Attraction?" *Semicond. Int.*, **22**, 46 (1999).
24. M. T. Bohr, "Interconnect Scaling—The Real Limiter to High Performance ULSI," *IEEE Tech. Dig. Int. Electron Devices Meet.*, p. 241, 1995.
25. E. Leobandung et al., "Scalability of SOI Technology into 0.13 μm 1.2 V CMOS Generation," *IEEE Int. Electron. Devices Meet.*, p. 403, 1998.
26. B. Martin, "Electronic Design Automation," *IEEE Spectr.*, **36**, 61 (1999).
27. K. Banerjee et al., "3-D ICs: A Novel Chip Design for Improving Deep-Submicrometer Interconnect Performance and Systems-on-Chip Integration", *Proc. IEEE*, **89**, 602 (2001).
28. H. Ishiuchi et al., "Embedded DRAM Technologies," *IEEE Tech. Dig. Int. Electron. Devices Meet.*, p. 33, 1997.
29. S. Luryi, J. Xu, and A. Zaslavsky, Eds, *Future Trends in Microelectronics*, Wiley, New York, 1999.

► **PROBLEMS (* DENOTES DIFFICULT PROBLEMS)**

FOR SECTION 15.1 PASSIVE COMPONENTS

1. For a sheet resistance of $1 \text{ k}\Omega/\square$, find the maximum resistance that can be fabricated on a $2.5 \times 2.5\text{-mm}$ chip using $2 \mu\text{m}$ lines with a $4 \mu\text{m}$ pitch (distance between the centers of the parallel lines).
2. Design a mask set for a 5 pF MOS capacitor. The oxide thickness is 30 nm . Assume that the minimum window size is $2 \times 10 \mu\text{m}$ and the maximum registration errors are $2 \mu\text{m}$.
3. Draw a complete step-by-step set of masks for the spiral inductor with three turns on a substrate.
4. Design a 10 nH square spiral inductor in which the total length of the interconnect is $350 \mu\text{m}$; the spacing between turns is $2 \mu\text{m}$.

FOR SECTION 15.2 BIPOLAR TECHNOLOGY

5. Draw the circuit diagram and device cross section of a clamped transistor.
6. Identify the purpose of the following steps in self-aligned double-polysilicon bipolar structure: (a) undoped polysilicon in trench in Fig. 12a, (b) the poly 1 in Fig. 12b, and (c) the poly 2 in Fig. 12d.

FOR SECTION 15.3 MOSFET TECHNOLOGY

- *7. In NMOS processing, the starting material is a p -type $10 \Omega\text{-cm}$ $\langle 100 \rangle$ -oriented silicon wafer. The source and drain are formed by arsenic implantation of $10^{16} \text{ ions/cm}^2$ at 30 keV through a gate oxide of 25 nm . (a) Estimate the threshold voltage change of the device. (b) Draw the doping profile along a coordinate perpendicular to the surface and passing through the channel region or the source region.
8. (a) Why is $\langle 100 \rangle$ -orientation preferred in NMOS fabrication? (b) What are the disadvantages if too thin a field oxide is used in NMOS devices? (c) What problems occur if a polysilicon gate is used for gate lengths less than $3 \mu\text{m}$? Can another material be substituted for polysilicon? (d) How is a self-aligned gate obtained and what are its advantages? (e) What purpose does P-glass serve?

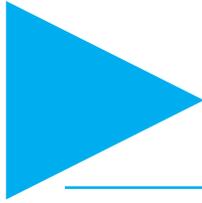
- *9. For a floating-gate nonvolatile memory, the lower insulator has a dielectric constant of 4 and is 10 nm thick. The insulator above the floating gate has a dielectric constant of 10 and is 100 nm thick. If the current density J in the lower insulators is given by $J = \sigma E$, where $\sigma = 10^{-7}$ S/cm, and the current in the other insulator is negligibly small, find the threshold voltage shift of the device caused by a voltage of 10 V applied to the control gate for (a) 0.25 μ s, and (b) a sufficiently long time that J in the lower insulator becomes negligibly small.
10. Draw a complete step-by-step set of masks for the CMOS inverter shown in Fig. 17. Pay particular attention to the cross section shown in Fig. 17c for your scale.
- *11. A 0.5 μ m digital CMOS technology has 5 μ m wide transistors. The minimum wire width is 1 μ m and the metallization layer consists of 1 μ m thick aluminum. Assume that μ_n is 400 cm²/V-s, d is 10-nm, V_{DD} is 3.3 V, and the threshold voltage is 0.6 V. Finally, assume that the maximum voltage drop that can be tolerated is 0.1 V when a 1 μ m² cross section aluminum wire is carrying the maximum current that can be supplied by the NMOS transistor. How long a wire can be allowed? Use a simple square-law, long-channel model to predict the MOS current drive (resistivity of aluminum is 2.7×10^{-8} Ω -cm).
12. Plot the cross-sectional views of a twin-tub CMOS structure of the following stages of processing: (a) n -tub implant, (b) p -tub implant, (c) twin-tub drive-in, (d) nonselective p^+ -source/drain implant, (e) selective n^+ -source/drain implant using photoresist as a mask, and (f) P-glass deposition.
13. Why do we use a p^+ -polysilicon gate for PMOS?
14. What is the boron penetration problem in p^+ -polysilicon PMOS? How would you eliminate it?
15. To obtain a good interfacial property, a buffered layer is usually deposited between the high- k material and substrate. Calculate the effective oxide thickness if the stacked gate dielectric structure is (a) a buffered nitride of 0.5 nm and (b) a Ta₂O₅ of 10 nm.
16. Describe the disadvantages of LOCOS technology and the advantages of shallow-trench isolation technology.

FOR SECTION 15.4 MESFET TECHNOLOGY

17. What is the purpose for the polyimide used in Fig. 31f?
18. Why is it difficult to make bipolar transistor and MOSFET in GaAs?

FOR SECTION 15.5 CHALLENGES FOR MICROELECTRONICS

19. Calculate the RC time constant of a aluminum runner 0.5 μ m thick formed on a thermally grown SiO₂ 0.5 μ m thick. The length and width of the runner are 1 cm and 1 μ m, respectively. The resistivity of the runner is 10^{-5} Ω -cm. (b) What will be the RC time constant for a polysilicon runner ($R_{\square} = 30$ Ω/\square) of identical dimension?
20. Why do we need multiple oxide thicknesses for a system-on-a-chip (SOC)?
21. Normally we need a buffered layer placed between a high- k Ta₂O₅ and the silicon substrate. Calculate the effective oxide thickness (EOT) when the stacked gate dielectric is Ta₂O₅ ($k = 25$) with a thickness of 75 \AA on a buffered nitride layer ($k = 7$ and a thickness of 10 \AA). Also, calculate EOT for a buffered oxide layer ($k = 3.9$, and a thickness of 5 \AA).



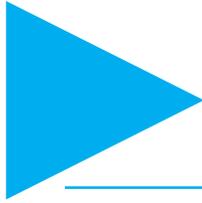
Appendix A

List of Symbols

| Symbol | Description | Unit |
|-----------------|-----------------------------------|--------------------|
| a | Lattice constant | Å |
| \mathcal{B} | Magnetic induction | Wb/m ² |
| c | Speed of light in vacuum | cm/s |
| C | Capacitance | F |
| \mathcal{D} | Electric displacement | C/cm ² |
| D | Diffusion coefficient | cm ² /s |
| E | Energy | eV |
| E_C | Bottom of conduction band | eV |
| E_F | Fermi energy level | eV |
| E_g | Energy bandgap | eV |
| E_V | Top of valence band | eV |
| \mathcal{E} | Electric field | V/cm |
| \mathcal{E}_c | Critical field | V/cm |
| \mathcal{E}_m | Maximum field | V/cm |
| f | Frequency | Hz(cps) |
| $F(E)$ | Fermi-Dirac distribution function | |
| h | Planck constant | J·s |
| $h\nu$ | Photon energy | eV |
| I | Current | A |
| I_C | Collector current | A |
| J | Current density | A/cm ² |
| J_{th} | Threshold current density | A/cm ² |
| k | Boltzmann constant | J/K |
| kT | Thermal energy | eV |
| L | Length | cm or μm |
| m_0 | Electron rest mass | kg |
| m_n | Electron effective mass | kg |
| m_p | Hole effective mass | kg |
| \bar{n} | Refractive index | |
| n | Density of free electrons | cm ⁻³ |
| n_i | Intrinsic carrier concentration | cm ⁻³ |

(continued)

| Symbol | Description | Unit |
|---|--|-------------------------------------|
| N | Doping concentration | cm^{-3} |
| N_A | Acceptor-impurity density | cm^{-3} |
| N_C | Effective density of states in conduction band | cm^{-3} |
| N_D | Donor-impurity density | cm^{-3} |
| N_V | Effective density of states in valence band | cm^{-3} |
| p | Density of free holes | cm^{-3} |
| P | Pressure | Pa |
| q | Magnitude of electronic charge | C |
| Q_{it} | Interface-trapped charge | charges/ cm^2 |
| R | Resistance | Ω |
| \mathcal{R} | Responsivity | A/W |
| t | Time | s |
| T | Absolute temperature | K |
| v | Carrier velocity | cm/s |
| v_s | Saturation velocity | cm/s |
| v_{th} | Thermal velocity | cm/s |
| V | Voltage | V |
| V_{bi} | Built-in potential | V |
| V_{EB} | Emitter-base voltage | V |
| V_B | Breakdown voltage | V |
| W | Thickness | cm or μm |
| W_B | Base thickness | cm or μm |
| ϵ_0 | Permittivity in vacuum | F/cm |
| ϵ_s | Semiconductor permittivity | F/cm |
| ϵ_{ox} | Insulator permittivity | F/cm |
| ϵ_s/ϵ_0 or ϵ_{ox}/ϵ_0 | Dielectric constant | |
| τ | Lifetime or decay time | s |
| θ | Angle | rad |
| λ | Wavelength | μm or nm |
| ν | Frequency of light | Hz |
| μ_0 | Permeability in vacuum | H/cm |
| μ_n | Electron mobility | $\text{cm}^2/\text{V}\cdot\text{s}$ |
| μ_p | Hole mobility | $\text{cm}^2/\text{V}\cdot\text{s}$ |
| ρ | Resistivity | $\Omega\text{-cm}$ |
| ϕ_{Bn} | Schottky barrier height on n -type semiconductor | V |
| ϕ_{Bp} | Schottky barrier height on p -type semiconductor | V |
| $q\phi_m$ | Metal work function | eV |
| ω | Angular frequency ($2\pi f$ or $2\pi\nu$) | Hz |
| $\overline{\omega}$ | Phonon frequency | eV |
| Ω | Ohm | Ω |

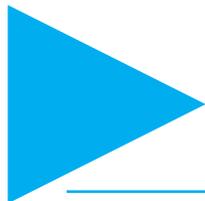


Appendix B

International System of Units (SI Units)

| Quantity | Unit | Symbol | Dimensions |
|---------------------|----------|--------|---------------------|
| Length [§] | meter | m | |
| Mass | kilogram | kg | |
| Time | second | s | |
| Temperature | kelvin | K | |
| Current | ampere | A | |
| Light intensity | candela | Cd | |
| Angle | radian | rad | |
| Frequency | hertz | Hz | 1/s |
| Force | newton | N | kg·m/s ² |
| Pressure | pascal | Pa | N/m ² |
| Energy [§] | joule | J | N·m |
| Power | watt | W | J/s |
| Electric charge | coulomb | C | A·s |
| Potential | volt | V | J/C |
| Conductance | siemens | S | A/V |
| Resistance | ohm | Ω | V/A |
| Capacitance | farad | F | C/V |
| Magnetic flux | weber | Wb | V·s |
| Magnetic induction | tesla | T | Wb/m ² |
| Inductance | henry | H | Wb/A |
| Light flux | lumen | Lm | Cd·rad |

[§] It is more common in the semiconductor field to use cm for length and eV for energy (1 cm = 10⁻² m, 1 eV = 1.6 × 10⁻¹⁹ J).

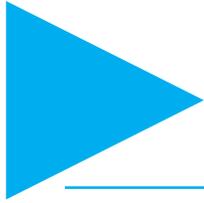


Appendix C

Unit Prefixes*

| Multiple | Prefix | Symbol |
|------------|--------|--------|
| 10^{18} | exa | E |
| 10^{15} | peta | P |
| 10^{12} | tera | T |
| 10^9 | giga | G |
| 10^6 | mega | M |
| 10^3 | kilo | k |
| 10^2 | hecto | h |
| 10 | deka | da |
| 10^{-1} | deci | d |
| 10^{-2} | centi | c |
| 10^{-3} | milli | m |
| 10^{-6} | micro | μ |
| 10^{-9} | nano | n |
| 10^{-12} | pico | p |
| 10^{-15} | femto | f |
| 10^{-18} | atto | a |

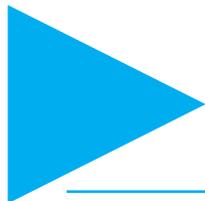
* As adopted by International Committee on Weights and Measures. (Compound prefixes should not be used, e.g., not $\mu\mu$ but p.)



Appendix D

Greek Alphabet

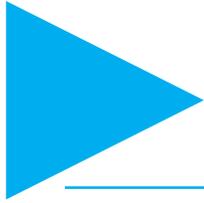
| Letter | Lowercase | Uppercase |
|---------|------------|-----------|
| Alpha | α | A |
| Beta | β | B |
| Gamma | γ | Γ |
| Delta | δ | Δ |
| Epsilon | ϵ | E |
| Zeta | ζ | Z |
| Eta | η | H |
| Theta | θ | Θ |
| Iota | ι | I |
| Kappa | κ | K |
| Lambda | λ | Λ |
| Mu | μ | M |
| Nu | ν | N |
| Xi | ξ | Ξ |
| Omicron | o | O |
| Pi | π | Π |
| Rho | ρ | P |
| Sigma | σ | Σ |
| Tau | τ | T |
| Upsilon | υ | Υ |
| Phi | ϕ | Φ |
| Chi | χ | X |
| Psi | ψ | Ψ |
| Omega | ω | Ω |



Appendix E

Physical Constants

| Quantity | Symbol | Value |
|----------------------------|--------------|--|
| Angstrom unit | Å | $10 \text{ Å} = 1 \text{ nm} = 10^{-3} \mu\text{m} = 10^{-7} \text{ cm} = 10^{-9} \text{ m}$ |
| Avogadro's number | N_{av} | 6.02214×10^{23} |
| Bohr radius | a_B | 0.52917 Å |
| Boltzmann constant | k | $1.38066 \times 10^{-23} \text{ J/K (R/N}_{av}\text{)}$ |
| Elementary charge | q | $1.60218 \times 10^{-19} \text{ C}$ |
| Electron rest mass | m_0 | $0.91094 \times 10^{-30} \text{ kg}$ |
| Electron volt | eV | $1 \text{ eV} = 1.60218 \times 10^{-19} \text{ J}$ $= 23.053 \text{ kcal/mol}$ |
| Gas constant | R | $1.98719 \text{ cal/mol-K}$ |
| Permeability in vacuum | μ_0 | $1.25664 \times 10^{-8} \text{ H/cm (} 4\pi \times 10^{-9}\text{)}$ |
| Permittivity in vacuum | ϵ_0 | $8.85418 \times 10^{-14} \text{ F/cm (} 1/\mu_0 c^2\text{)}$ |
| Planck constant | h | $6.62607 \times 10^{-34} \text{ J}\cdot\text{s}$ |
| Reduced Planck constant | \hbar | $1.05457 \times 10^{-34} \text{ J}\cdot\text{s (} h/2\pi\text{)}$ |
| Proton rest mass | M_p | $1.67262 \times 10^{-27} \text{ kg}$ |
| Speed of light in vacuum | c | $2.99792 \times 10^{10} \text{ cm/s}$ |
| Standard atmosphere | | $1.01325 \times 10^5 \text{ Pa}$ |
| Thermal voltage at 300 K | kT/q | 0.025852 V |
| Wavelength of 1 eV quantum | λ | $1.23984 \mu\text{m}$ |



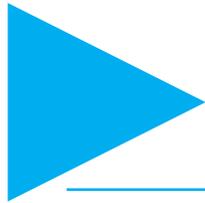
Appendix F

Properties of Important Element and Binary Compound Semiconductors at 300 K

| Semiconductor | Lattice constant (Å) | Bandgap (eV) | Band ^a | Mobility ^b (μ_n) | Mobility ^b (μ_p) | Dielectric constant |
|---------------|----------------------|--------------|-------------------|-----------------------------------|-----------------------------------|---------------------|
| Element | | | | | | |
| Ge | 5.65 | 0.66 | I | 3900 | 1800 | 16.2 |
| Si | 5.43 | 1.12 | I | 1450 | 505 | 11.9 |
| IV-IV | | | | | | |
| SiC | 3.08 | 2.86 | I | 300 | 40 | 9.66 |
| III-V | | | | | | |
| AlSb | 6.13 | 1.61 | I | 200 | 400 | 12.0 |
| GaAs | 5.65 | 1.42 | D | 9200 | 320 | 12.4 |
| GaP | 5.45 | 2.27 | I | 160 | 135 | 11.1 |
| GaSb | 6.09 | 0.75 | D | 3750 | 680 | 15.7 |
| InAs | 6.05 | 0.35 | D | 33000 | 450 | 15.1 |
| InP | 5.86 | 1.34 | D | 5900 | 150 | 12.6 |
| InSb | 6.47 | 0.17 | D | 77000 | 850 | 16.8 |
| II-IV | | | | | | |
| CdS | 5.83 | 2.42 | D | 340 | 50 | 5.4 |
| CdTe | 6.48 | 1.56 | D | 1050 | 100 | 10.2 |
| ZnO | 4.58 | 3.35 | D | 200 | 180 | 9.0 |
| ZnS | 5.42 | 3.68 | D | 180 | 10 | 8.9 |
| IV-VI | | | | | | |
| PbS | 5.93 | 0.41 | I | 800 | 1000 | 17.0 |
| PbTe | 6.46 | 0.31 | I | 600 | 4000 | 30.0 |

^aI, indirect, D, direct.

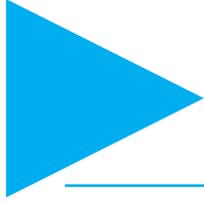
^bThe values are for drift mobilities obtained in the purest and most perfect materials available to date.



Appendix G

Properties of Si and GaAs at 300 K

| Properties | Si | GaAs |
|--|--|---|
| Atoms/cm ³ | 5.02×10^{22} | 4.42×10^{22} |
| Atomic weight | 28.09 | 144.63 |
| Breakdown field (V/cm) | $\sim 3 \times 10^5$ | $\sim 4 \times 10^5$ |
| Crystal structure | Diamond | Zincblende |
| Density (g/cm ³) | 2.329 | 5.317 |
| Dielectric constant | 11.9 | 12.4 |
| Effective density of states in conduction band, N_C (cm ⁻³) | 2.86×10^{19} | 4.7×10^{17} |
| Effective density of states in valence band, N_V (cm ⁻³) | 2.66×10^{19} | 7.0×10^{18} |
| Effective mass (conductivity) | | |
| Electrons (m_n/m_0) | 0.26 | 0.063 |
| Holes (m_p/m_0) | 0.69 | 0.57 |
| Electron affinity, χ (V) | 4.05 | 4.07 |
| Energy gap (eV) | 1.12 | 1.42 |
| Index of refraction | 3.42 | 3.3 |
| Intrinsic carrier concentration (cm ⁻³) | 9.65×10^9 | 2.25×10^6 |
| Intrinsic resistivity (Ω -cm) | 3.3×10^5 | 2.9×10^8 |
| Lattice constant (\AA) | 5.43102 | 5.65325 |
| Linear coefficient of thermal expansion, $\Delta L/L \times T$ ($^{\circ}\text{C}^{-1}$) | 2.59×10^{-6} | 5.75×10^{-6} |
| Melting point ($^{\circ}\text{C}$) | 1412 | 1240 |
| Minority-carrier lifetime (s) | 3×10^{-2} | $\sim 10^{-8}$ |
| Mobility (cm ² /V-s) | | |
| μ_n (electrons) | 1450 | 9200 |
| μ_p (holes) | 505 | 320 |
| Specific heat (J/g- $^{\circ}\text{C}$) | 0.7 | 0.35 |
| Thermal conductivity (W/cm-K) | 1.31 | 0.46 |
| Vapor pressure (Pa) | 1 at 1650 $^{\circ}\text{C}$ 10-6 at 900 $^{\circ}\text{C}$ | 100 at 1050 $^{\circ}\text{C}$ 1 at 900 $^{\circ}\text{C}$ |



Appendix H

Derivation of the Density of States in a Semiconductor

3-D Density of States

For a three dimensional (3-D) structure such as a bulk semiconductor, to calculate the electron and hole concentrations in the conduction and valence bands, respectively, we need to know the density of states, that is, the number of allowed energy states per unit energy per unit volume (i.e., in the unit of number of states/eV/cm³).

When electrons move back and forth along the x -direction in a semiconductor material, the movements can be described by standing-wave oscillations. The wavelength λ of a standing wave is related to the length of the semiconductor L by

$$\frac{L}{\lambda} = n_x, \quad (1)$$

where n_x is an integer. The wavelength can be expressed by de Broglie hypothesis:

$$\lambda = \frac{h}{p_x}, \quad (2)$$

where h is the Planck's constant and p_x is the momentum in the x -direction. Substituting Eq. 2 into Eq. 1 gives

$$Lp_x = hn_x. \quad (3)$$

The incremental momentum dp_x required for a unity increase in n_x is

$$Ldp_x = h. \quad (4)$$

For a three-dimensional cube of side L , we have

$$L^3 dp_x dp_y dp_z = h^3. \quad (5)$$

The volume $dp_x dp_y dp_z$ in the momentum space for a unit cube ($L = 1$) is thus equal to h^3 . Each incremental change in n corresponds to a unique set of integers (n_x, n_y, n_z), which in turn corresponds to an allowed energy state. Thus, the volume in momentum space for an energy state is h^3 . Figure 1 shows the momentum space in spherical coordinates. The volume between two concentric spheres (from p to $p+dp$) is $4\pi p^2 dp$. The number of energy states contained in the volume is then $2(4\pi p^2 dp)/h^3$, where the factor 2 accounts for the electron spins.

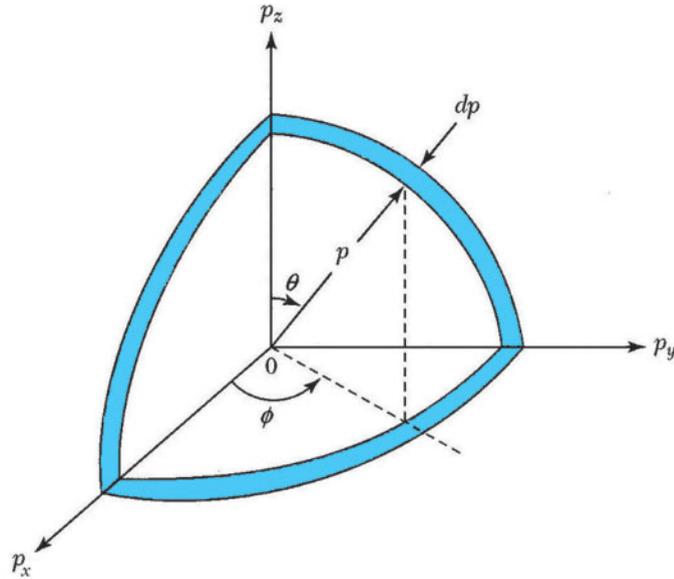


Fig. 1 The momentum space in spherical coordinates.

The energy E of the electron (here we consider only the kinetic energy) is given by

$$E = \frac{p^2}{2m_n} \quad (6)$$

or

$$p = \sqrt{2m_n E} \quad (7)$$

where p is the total momentum (with components p_x, p_y and p_z in Cartesian coordinates) and m_n is the effective mass. From Eq. 7, we can substitute E for p and obtain

$$N(E)dE = \frac{8\pi p^2 dp}{h^3} = 4\pi \left(\frac{2m_n}{h^2} \right)^{\frac{3}{2}} E^{\frac{1}{2}} dE \quad (8)$$

and

$$N(E) = 4\pi \left(\frac{2m_n}{h^2} \right)^{\frac{3}{2}} E^{\frac{1}{2}}, \quad (9)$$

where $N(E)$ is called the density of states. The $N(E)$ varies with \sqrt{E} as shown in Fig. 2a.

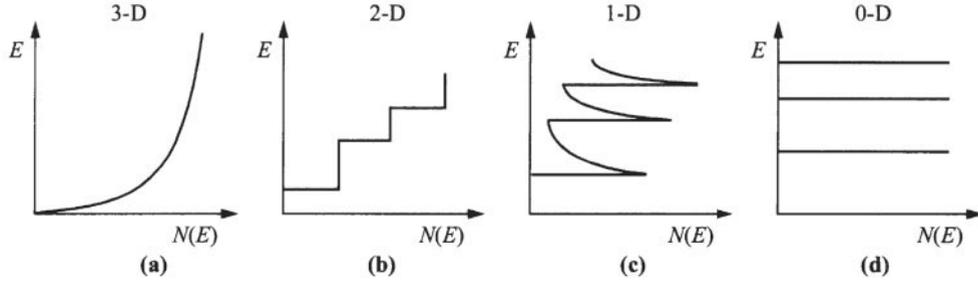


Fig. 2 Density of states $N(E)$ for (a) bulk semiconductor (3-D), (b) quantum well (2-D), (c) quantum wire (1-D), and (d) quantum dot (0-D)

2-D Density of States

In two-dimensional structures such as the quantum well, the derivation of 2-D density of states is almost the same as for 3-D except that one of the p -space components is fixed. Instead of finding the number of p -states enclosed within a sphere, we calculate the number of p -states lying in an annulus of radius p to $p + dp$. The incremental momentum dp_x required for a unity increase in n_x is

$$Ldp_x = h. \quad (10)$$

For a two-dimensional square of side L , we have

$$L^2 dp_x dp_y = h^2. \quad (11)$$

The area $dp_x dp_y$ in the momentum space for a unit square ($L = 1$) is thus equal to h^2 . Figure 3 shows the momentum space in circular coordinates. The area between two concentric circles (from p to $p+dp$) is $2\pi p dp$. The number of energy states contained in the area is then $2(2\pi p dp)/h^2$, where the factor 2 accounts for the electron spins.

$$N(E)dE = \frac{4\pi p dp}{h^2} = 4\pi \left(\frac{m_n}{h^2} \right) dE \quad (12)$$

$$N(E) = \frac{4\pi m_n}{h^2} = \frac{m_n}{\pi \hbar^2}. \quad (13)$$

The 2-D density of states does not depend on energy. As the top of the energy gap is reached, there is a significant number of available states. Taking into account the other energy levels in the quantum well, the density of states becomes the staircase-like function shown in Fig. 2b.

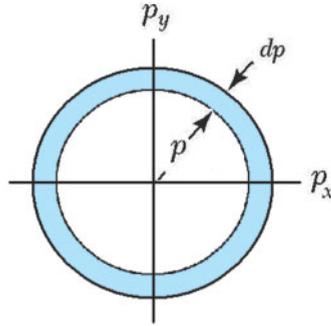


Fig. 3 The momentum space in circular coordinates.

1-D Density of States

In one-dimensional structures such as the quantum wire, two of the p -space components are fixed. Compared with 2-D, the p -space becomes a line. The wavelength λ of a standing wave is related to the length of the semiconductor L by

$$\frac{L}{\lambda/2} = n_x, \quad (14)$$

The incremental momentum dp_x required for a unity increase in n_x is

$$2Ldp_x = h, \quad (15)$$

The dp_x in the momentum space for a line with a unit length ($L=1$) is thus equal to $h/2$.

Figure 4 shows the momentum space in line coordinates. The length between p to $p+dp$ is dp . The number of energy states contained in the line is then $2dp/(h/2)$, where the factor 2 accounts for the electron spins.

$$N(E)dE = \frac{2dp}{h/2} = 2 \left(\frac{2m_n}{E} \right)^{1/2} \frac{1}{h} dE = \frac{1}{\pi} \left(\frac{2m_n}{\hbar^2} \right)^{1/2} \frac{1}{E^{1/2}} dE \quad (16)$$

$$N(E) = \frac{1}{\pi} \left(\frac{2m_n}{\hbar^2} \right)^{1/2} \frac{1}{E^{1/2}}. \quad (17)$$

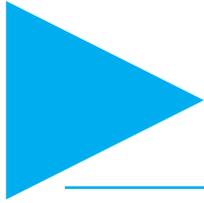
The $N(E)$ varies with $E^{-1/2}$ as shown in Fig. 2c.

0-D Density of States

In a 0-D structure such as a quantum dot, the values of p are quantized in all directions. All the available states exist only at discrete energies and can be represented by a delta function as shown in Fig. 2d. The density of states in a quantum dot is continuous and independent of energy. In a real quantum dot, however, the size distribution leads to a broadening of this line function.



Fig. 4 The momentum space in line coordinates.



Appendix I

Derivation of Recombination Rate for Indirect Recombination

A schematic diagram of the various transitions that occur in recombination through recombination centers is shown in Fig. 13 of Chapter 2. If the concentration of centers in the semiconductor is N_t , the concentration of unoccupied centers is given by $N_t(1-F)$, where F is the Fermi distribution function for the probability that a center is occupied by an electron. In equilibrium,

$$F = \frac{1}{1 + e^{(E_t - E_F)/kT}} \quad (1)$$

where E_t is the energy level of the center and E_F is the Fermi level.

Therefore, the capture rate of an electron by a recombination center (Fig. 13a of Chapter 2) is given by

$$R_a \approx nN_t(1 - F). \quad (2)$$

We designate the proportionality constant by the product $v_{th}\sigma_n$, so that

$$R_a = v_{th}\sigma_n nN_t(1 - F). \quad (3)$$

The product $v_{th}\sigma_n$ may be visualized as the volume swept out per unit time by an electron with cross section σ_n . If the center lies within this volume, the electron will be captured by it.

The rate of emission of electrons from the center (Fig. 13b) is the inverse of the electron capture process. The rate is proportional to the concentration of centers occupied by electrons, that is, $N_t F$. We have

$$R_b = e_n N_t F. \quad (4)$$

The proportionality constant e_n is called the emission probability. At thermal equilibrium the rates of capture and emission of electrons must be equal ($R_a = R_b$). Thus, the emission probability can be expressed in terms of the quantities already defined in Eq. 3:

$$e_n = \frac{v_{th}\sigma_n n(1 - F)}{F}. \quad (5)$$

Since the electron concentration in thermal equilibrium is given by

$$n = n_i e^{(E_F - E_i)/kT}, \quad (6)$$

we obtain

$$e_n = v_{th} \sigma_n n_i e^{(E_i - E_c)/kT} . \quad (7)$$

The transitions between the recombination center and valence band are analogous to those described above. The capture rate of a hole by an occupied recombination center (Fig. 13c) is given by

$$R_c = v_{th} \sigma_p p N_t F . \quad (8)$$

By arguments similar to those for electron emission, the rate of hole emission (Fig. 13d) is

$$R_d = e_p N_t (1 - F) . \quad (9)$$

The emission probability e_p of a hole may be expressed in terms of v_{th} and σ_p by considering the thermal equilibrium condition for which $R_c = R_d$:

$$e_p = v_{th} \sigma_p n_i e^{(E_i - E_v)/kT} . \quad (10)$$

Let us now consider the nonequilibrium case in which an n -type semiconductor is illuminated uniformly to give a generation rate G_L . Thus in addition to the process shown in Fig. 13, electron-hole pairs are generated as a result of light. In steady state the electrons entering and leaving the conduction band must be equal. This is called the *principle of detailed balance*, and it yields

$$\frac{dn_n}{dt} = G_L - (R_a - R_b) = 0 . \quad (11)$$

Similarly, in steady state the detailed balance of holes in valence band leads to

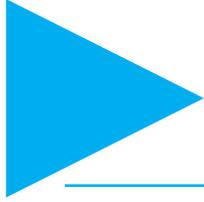
$$\frac{dp_p}{dt} = G_L - (R_c - R_d) = 0 . \quad (12)$$

Under equilibrium conditions, that is, $G_L = 0$, $R_a = R_b$, and $R_c = R_d$. However, under state-state nonequilibrium conditions, $R_a \neq R_b$, and $R_c \neq R_d$. From Eqs. 11 and 12 we obtain

$$G_L = R_a - R_b = R_c - R_d \equiv U . \quad (13)$$

We can get the net recombination rate U from Eqs. 3, 4, 8, and 9:

$$U \equiv R_a - R_b = \frac{v_{th} \sigma_n \sigma_p N_t (p_n n_n - n_i^2)}{\sigma_p [p_n + n_i e^{(E_i - E_c)/kT}] + \sigma_n [n_n + n_i e^{(E_i - E_v)/kT}]} . \quad (14)$$



Appendix J

Calculation of the Transmission Coefficient for a Symmetric Resonant-Tunneling Diode

To calculate the transmission coefficient, we consider Fig. 9a of Chapter 8 where the five regions (I, II, III, IV, V) are specified by the coordinates (x_1, x_2, x_3, x_4) . The Schrödinger equation for an electron in any region can be written as

$$\frac{\hbar^2}{2m_i^*} \left(\frac{d^2 \psi_i}{dx^2} \right) + V_i \psi_i = E \psi_i \quad i = 1, 2, 3, 4, 5, \quad (1)$$

where \hbar is the reduced Planck constant, m_i^* the effective mass in the i th region, E the incident energy, and V_i and ψ_i the potential energy and the wave function in the i th region, respectively. The wavefunction ψ_i can be expressed as

$$\psi_i(x) = A_i \exp(jk_i x) + B_i \exp(-jk_i x), \quad (2)$$

where A_i and B_i are constants to be determined from the boundary conditions and $k_i = \sqrt{2m_i^*(E - V_i)}/\hbar$. Since the wavefunctions and their first derivatives (i.e., $\psi_i / m_i^* = \psi_{i+1} / m_{i+1}^*$) at each potential discontinuity must be continuous, we obtain the transmission coefficient (for identical effective mass across the five regions)

$$T_t = \frac{1}{1 + E_0^2 (\sinh^2 \beta L_B) H^2 / [4E^2 (E_0 - E)^2]}, \quad (3)$$

where

$$H \equiv 2[E(E_0 - E)]^{1/2} \cosh \beta L_B \cos kL_W - (2E - E_0) \sinh \beta L_B \sin kL_W,$$

and

$$\beta \equiv \frac{\sqrt{2m^*(E_0 - E)}}{\hbar} \quad k = \frac{\sqrt{2m^*E}}{\hbar}$$

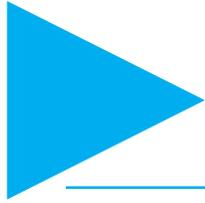
The resonant condition occurs when $H = 0$, and thus $T_t = 1$. The resonant-tunneling energy levels E_n can be calculated by solving the transcendental equation:

$$\frac{2[E(E_0 - E)]^{1/2}}{(2E - E_0)} = \tan kL_w \tanh \beta L_B, \quad (4)$$

As a first-order estimate of the energy levels, one can use the results of a quantum well with infinite barrier height:

$$E_n \approx \left(\frac{\pi^2 \hbar^2}{2m^* L_w^2}\right)n^2. \quad (5)$$

For a double-barrier structure with finite barrier height and width, the energy level (for a given n) will be lower; however, it will have a similar dependence on the effective mass and well width, that is, E_n increases with decreasing m^* or L_w .



Appendix K

Basic Kinetic Theory of Gases

The ideal gas law states that

$$PV = RT = N_{av}kT, \quad (1)$$

where P is the pressure, V is the volume of one mole of gas, R is the gas constant (1.98 cal/mol-K, or 82 atm-cm³/mol-K), T is the absolute temperature in K , N_{av} is Avogadro's number (6.02×10^{23} molecules/mole), and k is Boltzmann's constant (1.38×10^{-23} J/K, or 1.37×10^{-22} atm-cm²/K). Since real gases behave more and more like the ideal gas as the pressure is lowered, Eq. 1 is valid for most vacuum processes. We can use Eq. 1 to calculate the molecular concentration n (the number of molecules per unit volume):

$$n = \frac{N_{av}}{V} = \frac{P}{kT} \quad (2)$$

$$= 7.25 \times 10^{16} \left(\frac{P}{T} \right) \text{ molecules/cm}^3, \quad (2a)$$

where P is in Pa. The density ρ_d of a gas is given by the product of its molecular weight and its concentration:

$$\rho_d = \text{molecular weight} \times \left(\frac{P}{kT} \right) \quad (3)$$

The gas molecules are in constant motion and their velocities are temperature dependent. The distribution of velocities is described by the Maxwell-Boltzmann distribution law, which states that for a given speed v ,

$$\frac{1}{n} \frac{dn}{dv} \equiv f_v = \frac{4}{\sqrt{\pi}} \left(\frac{m}{2kT} \right)^{3/2} v^2 \exp\left(-\frac{mv^2}{2kT} \right), \quad (4)$$

where m is the mass of a molecule. This equation states that if there are n molecules in the volume, there will be dn molecules having a speed between v and $v + dv$. The average speed can be obtained from Eq. 4:

$$v_{av} = \frac{\int_0^{\infty} v f_v dv}{\int_0^{\infty} f_v dv} = \frac{2}{\sqrt{\pi}} \sqrt{\frac{2kT}{m}}. \quad (5)$$

An important parameter for vacuum technology is the molecular *impingement rate*, that is, how many molecules impinge on a unit area per unit time. To obtain this parameter, first consider the distribution function f_{v_x} for the velocities of molecules in the x -direction. This function can be expressed by an equation similar to Eq. 4:

$$\frac{1}{n} \frac{dn_x}{dv_x} \equiv f_{v_x} = \left(\frac{m}{2\pi kT}\right)^{1/2} v_x^2 \exp\left(\frac{-mv_x^2}{2kT}\right). \quad (6)$$

The molecular impingement rate ϕ is given by

$$\phi = \int_0^{\infty} v_x dn_x. \quad (7)$$

Substituting dn_x from Eq. 6 and integrating gives

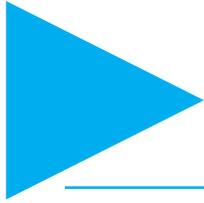
$$\phi = n \sqrt{\frac{kT}{2\pi m}}. \quad (8)$$

The relationship between the impingement rate and the gas pressure is obtained by using Eq. 2:

$$\phi = P(2\pi mkT)^{-1/2}. \quad (9)$$

$$= 2.64 \times 10^{20} \left(\frac{P}{MT}\right) \quad (9a)$$

where P is the pressure in Pa and M is the molecular weight.



Appendix L

Answers to Selected Problems

Answers are provided for those odd-numbered problems that have numerical solutions.

CHAPTER 1

- (a) 2.35 \AA ; (b) 6.78×10^{14} (100), 9.6×10^{15} (110), 7.83×10^{14} (111) atoms/cm².
- 52% (simple cubic), 74% (fcc), 34% (diamond).
- (643) plane
- E_{gSi} (100 K, 600 K) = 1.163, 1.032 eV, E_{gGaAs} (100 K, 600 K) = 1.501, 1.277 eV
- $m_{pSi} = 1.03 m_0$, $m_{pGaAs} = 0.43 m_0$.
- $KE = 3/2 kT$.
- $T = 1000/1.8 = 555 \text{ K}$ or $282 \text{ }^\circ\text{C}$.
- $p_0 = 9.3 \times 10^2 \text{ cm}^{-3}$; $E_F - E_i = 0.42 \text{ eV}$.
- $N_D = 2N_A$; $2N_A$.
- The ratio of $N_D^0/N_D^+ = 0.876$.

CHAPTER 2

- $3.31 \times 10^5 \text{ } \Omega\text{-cm}$ (Si), $2.92 \times 10^8 \text{ } \Omega\text{-cm}$ (GaAs).
- $167 \text{ cm}^2/\text{V}\cdot\text{s}$.
- $3.5 \times 10^7 \text{ cm}^{-3}$, $400 \text{ cm}^2/\text{V}\cdot\text{s}$.
- $0.226 \text{ } \Omega\text{-cm}$.
- $N_A = 50N_D$.
- (b) 259 V/cm .
- (a) $\Delta n = 10^{11} \text{ cm}^{-3}$, $n = 10^{15} \text{ cm}^{-3}$, $p = 10^{11} \text{ cm}^{-3}$
- $p(x) = 10^{14}(1 - 0.9e^{-x/L_p})$.
- 0.403, 7.8×10^{-9}
- $1.35 \times 10^5 \text{ cm/s} < 9.5 \times 10^6 \text{ cm/s}$ (100 V/cm)
 $1.35 \times 10^7 \text{ cm/s} \approx 9.5 \times 10^6 \text{ cm/s}$ (10^4 V/cm).

CHAPTER 3

- $1.867 \text{ } \mu\text{m}$, $V_{bi} = 0.52 \text{ V}$, $\mathcal{E}_m = 4.86 \times 10^3 \text{ V/cm}$.
- At 300 K, $V_{bi} = 0.714 \text{ V}$, $W = 0.97 \text{ } \mu\text{m}$
 $\mathcal{E}_m = 1.47 \times 10^4 \text{ V/cm}$.
- For $N_D = 10^{15} \text{ cm}^{-3}$,
 $1/C_j^2 = 1.187 \times 10^{16}$ (0.834-V).
- $N_D = 3.43 \times 10^{15} \text{ cm}^{-3}$.
- $2.5 \times 10^{17} \text{ cm}^{-3}$.
- $N_A = 2.2 \times 10^{15} \text{ cm}^{-3}$, $N_D = 5.4 \times 10^{15} \text{ cm}^{-3}$.
- 0.79 V
- $8.78 \times 10^{-3} \text{ C/cm}^2$.
- cross-sectional area is $8.6 \times 10^{-5} \text{ cm}^2$.
- (a) 587 V (b) 42.8 V
- For $V = +0.5 \text{ V}$, $V_{bi} = 1.1$ and $3.4 \times 10^{-4} \text{ V}$;
depletion widths = 3.82×10^{-5} and $1.27 \times 10^{-8} \text{ cm}$

CHAPTER 4

- (a) 0,995, 199, (b) $2 \times 10^{-6} \text{ A}$.
- (a) $0.904 \text{ } \mu\text{m}$, (b) $2.54 \times 10^{11} \text{ cm}^{-3}$.
- (a) $I_E = 1.606 \times 10^{-5} \text{ A}$, $I_C = 1.596 \times 10^{-5} \text{ A}$
 $I_B = 1.041 \times 10^{-7} \text{ A}$; (b) $\beta_0 = 160$.
- $D_E = 2.269 \text{ cm/s}$, $D_B = 30.73 \text{ cm/s}$,
 $D_C = 11.75 \text{ cm/s}$.
- $\beta_0 = 50,000$.
- 131.6.
- $I_E = 1.715 \times 10^{-4} \text{ A}$, $I_C = 1.715 \times 10^{-4} \text{ A}$
- $f_T = 1.27 \text{ GHz}$, $f_\alpha = 1,275 \text{ GHz}$,
 $f_\beta = 2.55 \text{ MHz}$.

25. 0.29.
27. 3.96×10^{-3} cm, 44.4 cm².

CHAPTER 5

5. 0.15 μ m.
7. 0.59 V, 1.11×10^5 V/cm.
9. 2.32×10^{-2} V.
11. 7.74×10^{-2} V.
15. 3.42 V, 2.55×10^{-2} A.
17. 3.45×10^{-4} S,
19. 8×10^{11} cm⁻².
21. 1.7×10^{12} cm⁻².
23. 0.457 μ m.
25. 0.83 V.

CHAPTER 6

1. (a) $1/2000$. (b) 1 mJ.
7. 0.18 V.
9. (a) It is fully depleted with a SOI structure, the leakage current is low. (b) Source and drain to body junction capacitance is low, the speed is fast. (c) The Si is very thin, there is no leakage path far from the gate. (d) Low vertical field and less impurity scattering, the mobility is higher.
11. 0.46 μ m.
13. 9 V.
15. 8094 electrons.
17. 4.34 V.

CHAPTER 7

1. 0.54 eV, $V_{bi} = 0.352$ V.
3. $\phi_{Bn} = 0.64$ V, $V_{bi} = 0.463$ V, $W = 0.142$ μ m,
 $\mathcal{E}_m = 6.54 \times 10^4$ V/cm.
5. $V_{bi} = 0.605$ V, $\phi_m = 4.81$ V.
7. 0.108 μ m.
9. (b) -2.06 V.
11. 0.152 μ m, 0.0496 μ m.
13. 5.8 nm.
15. 44.5 nm, -0.93 V.

CHAPTER 8

1. 18.9 nm, 6.13×10^{-7} F/cm².
3. (a) 137 Ω ; (b) 318.5 V.
5. (a) 74.8 V; (b) 2.2×10^5 V/cm; (c) 19 GHz
7. (a) 10^{16} cm⁻³; (b) 10 ps; (c) 2.02 W.
9. 3 meV, 11 meV.

CHAPTER 9

1. 1.57 mW.
3. 18 nm.
5. 1.6 GHz.
7. 0.70 .
9. 9.5 mW.
11. 8.83% , 5.95% .
13. 0.828 nm, 147 GHz.
15. 0.794 , 0.998 .
17. $g(84^\circ) = 270$, $g(78^\circ) = 217$, 50.43 μ m.
19. $\lambda_B = 1.3296$ or 1.3304 mm, $\Lambda = 0.196$ μ m.

CHAPTER 10

1. (1) 0.645 A/W, (2) 0 , (3) 0.645 A/W, (4) 0 .
3. $I_p = 2.55$ μ A, gain = 9 .
7. 138 V, 7×10^{15} cm⁻³, 90 ps.
9. 5 μ m.
13. $P_{load} = 0.032$ W, $\eta = 13\%$.
15. Maximum power output = 1.49 W, $FF = 0.83$.
17. For amorphous Si, 0.23 μ m, CIGS, 2.3 μ m.

CHAPTER 11

1. At $x = 0$, $C_s = 3 \times 10^{16}$ cm⁻³; at $x = 0.9$,
 $C_s = 1.5 \times 10^{17}$ cm⁻³.
3. 0.75 g.
5. 6.56 m.
9. 24 cm.
11. $\pm 30\%$, $\pm 1\%$.
15. 2.14×10^{14} cm⁻³ at 900 °C.
17. 4.68×10^4 cm/s.
19. 5.27×10^{14} atoms/cm².
21. 0.25

CHAPTER 12

1. 44 min.
5. (a) $x = 0.83$, $y = 0.46$; (b) $2 \times 10^{11} \Omega\text{-cm}$
7. 0.0093.
9. 757 °C.
13. 7×10^{14} molecules/cm².
15. 2.1×10^{11} cm⁻².
17. 71.1 nm for TiSi₂.
19. (a) 0.93 ns; (b) 0.42 ns; (c) 0.45.
21. 72 Ω, 0.18 V.

CHAPTER 13

1. (a) 2765; (b) 578; (c) 157.
3. 7 wafers/hr for positive resist, 120 for negative resist.
9. (a) $W_b = 1.22 \mu\text{m}$; (b) 0.93 μm; (c) 0.65 μm.
11. Etch from the top, wafer area lost = 127 cm²; etch from the bottom, wafer area lost is small.
13. 224.7 nm/min.
17. 433.3 nm.

CHAPTER 14

1. 0.15 μm, 5.54×10^{14} atoms/cm².
3. 25 min, 3.4×10^{13} atoms/cm².
5. 16.9%.
7. $x_j = 32.3$ nm.
11. 6.7 s.
13. 0.53 μm.
15. 99.6%.
21. 0.927 μm.

CHAPTER 15

1. 781 MΩ.
3. 13 turns.
7. (a) 0.91 V; (b) peak concentration = 2.2×10^{21} cm⁻³.
9. (a) 0.565 V; (b) 9.98 V.
11. 740 μm.
15. 1.84 nm.
19. 1.38 ns, 207 ns.
21. 17.3 Å, 16.7 Å.

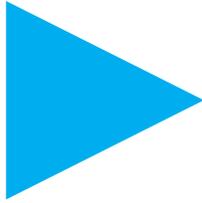


Photo Credits

▶ CHAPTER 0

Figure 0.3 Reprinted with permission of Alcatel-Lucent USA Inc.

Figure 0.4 Reprinted with permission of Alcatel-Lucent USA Inc.

Figure 0.6 Courtesy Dr. Gordon Moore

Figure 0.7 Courtesy Intel Corp

▶ CHAPTER 6

Figure 6.18 © 2007 IEEE. Reprinted, with permission, from IEEE Proceedings, A. Theuwissen, "CMOS image sensors: State-of-the-art and future perspectives."

Figure 6.19 © 2007 IEEE. Reprinted, with permission, from IEEE Proceedings, A. Theuwissen, "CMOS image sensors: State-of-the-art and future perspectives."

Figure 6.31 © 1984 IEEE. Reprinted, with permission, from IEEE Proceedings, F. Masuoka et. al, "A new flash E2PROM cell using triple polysilicon technology."

▶ CHAPTER 10

Figure 10.9 Reprinted with permission from Applied Physics Letters, 40 (38), F. Capasso et. al, "Enchantment of Electron Impact Ionisation in a Superlattice: A New Avalanche Photodiode with a Large Ionisation Rate Ratio", Copyright 1982 American Institute of Physics.

Figure 10.21 Reprinted from Solar Energy Materials and Solar Cells, 78, A. V. Shah et. al, "Material and solar cell research in microcrystalline silicon", pp. 469-491, Copyright 2003, with permission from Elsevier.

Figure 10.22 © 1980 IEEE. Reprinted, with permission, from IEEE Transactions on Electron Devices, A.M. Barnett and A. Rothwarf, "Thin-film solar cells: A unified analysis of their potential."

Figure 10.24 Reprinted with permission from Applied Physics Letters, 95 (6), Ta-Ya Chu et. al, "Highly efficient polycarbazole-based organic photovoltaic devices", Copyright 2009, American Institute of Physics.

Figure 10.26 Reprinted from Physica E: Low-Dimensional Systems and Nanostructures, 14, A.J. Nozik, "Quantum dot solar cells", pp. 115-120, Copyright 2002, with permission from Elsevier.

▶ CHAPTER 11

Figure 11.13 Photograph courtesy of Shin-Etsu Chemical Co, Ltd.

Figure 11.23 M.B. Panish, Bell Laboratories, Lucent Technologies.

Figure 11.27 Reprinted with permission from Journal of Applied Physics, 68 (7), S.F. Fang et. al, "Gallium arsenide and other compound semiconductors on silicon", Copyright 1990, American Institute of Physics.

▶ CHAPTER 12

Figure 12.17 Handbook of Semiconductor Manufacturing Technology by Robert Doering. Copyright 2007. Reproduced with permission of Taylor & Francis Group LLC - Books in the format textbook via Copyright Clearance Center

Figure 12.18 Handbook of Semiconductor Manufacturing Technology by Robert Doering. Copyright 2007. Reproduced with permission of Taylor & Francis Group LLC - Books in the format textbook via Copyright Clearance Center

▶ CHAPTER 13

Figure 13.22 Handbook of Semiconductor Manufacturing Technology by Robert Doering. Copyright 2007. Reproduced with permission of Taylor & Francis Group LLC - Books in the format textbook via Copyright Clearance Center

Figure 13.25 Handbook of Semiconductor Manufacturing Technology by Robert Doering. Copyright 2007. Reproduced with permission of Taylor & Francis Group LLC - Books in the format textbook via Copyright Clearance Center

Figure 13.27 Handbook of Semiconductor Manufacturing Technology by Robert Doering. Copyright 2007. Reproduced with permission of Taylor & Francis Group LLC - Books in the format textbook via Copyright Clearance Center

▶ CHAPTER 14

Figure 14.24 Handbook of Semiconductor Manufacturing Technology by Robert Doering. Copyright 2007. Reproduced with permission of Taylor & Francis Group LLC - Books in the format textbook via Copyright Clearance Center

▶ CHAPTER 15

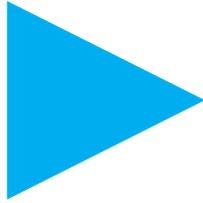
Figure 15.22 © 2010 IEEE. Reprinted, with permission, from IEEE Proceedings, J. Howard et. al, "A 48-Core IA-32 message-passing processor with DVFS in 45nm CMOS."

Figure 15.24 © 2000 IEEE. Reprinted, with permission, from IEEE Transactions on Electron Devices, D. Hisamoto et. al, "FinFET-a self-aligned double-gate MOSFET scalable to 20 nm."

Figure 15.25 © 2000 IEEE. Reprinted, with permission, from IEEE Transactions on Electron Devices, D. Hisamoto et. al, "FinFET-a self-aligned double-gate MOSFET scalable to 20 nm."

Figure 15.28 © 2009 IEEE. Reprinted, with permission, from IEEE Proceedings, U. Kang et. al, "8Gb 3D DDR3 DRAM using through-silicon-via technology."

Figure 15.29 © 2009 IEEE. Reprinted, with permission, from IEEE Proceedings, C. Trinh et. al, "A 5.6 MB/s 64Gb 4b/Cell NAND Flash memory in 43nm CMOS."



Index

- <100> orientation, 170, 172,
- <100> plane, 22, 397, 448, 464
- <100> direction, 492
- <100> plane, 22, 41, 305, 448
- <100> direction, 488
- <111>-axis, 488
- <111> direction, 22, 41
- <111> orientation, 170
- <111> plane, 343, 397, 448, 449, 464
- <111> surface, 170
- 3-D integration, 7, 215
 - system, 9
- 48-core microprocessor, 507, 524
- 5.6 MB/s, 64 Gb, 4b/cell NAND flash memory, 529, 538
- 8 Gb DRAM, 528

- Abrupt heterojunction, 118
- Abrupt junction, 86-91, 94, 96, 98, 108, 112-117, 121-122
- Absorption, 281-285, 259, 282, 328
 - coefficient, 284, 285, 289, 321, 322, 326, 330, 334, 347, 348, 355, 356
- Acceptor, 15, 34, 35, 36, 41, 42
- Acceptor states, 170, 171
- Acceptor vacancy, 476, 478
- Accumulation, 162, 163, 166, 192, 203
- Acetic acid, 447-449
- Activation energy, 108, 397, 403, 406, 410, 420, 469, 479, 492
- Active mode, 125-128, 130-132, 134-137, 141-142, 144, 147, 158
- Air mass 336, 337
- Air mass 0, 337
- Air mass 1.5, 337
- Air molecules, 383
- Air traffic control beacons, 259
- AlCu etch chamber, 459
- AlGaAs, 249, 251, 252, 254, 257, 311-314, 322, 275, 335, 343
- AlGaAs/GaAs heterojunction, 252, 257
- AlGaInN, 293
- AlInAs/GaInAs HBT, 274
- AlN, 293
- Al₂O₃, 175, 176, 201, 409, 414, 417, 421, 427
- AlSb, 537
- Aluminum (Al) 15, 18, 19, 169, 202, 221, 239, 345, 297, 303, 344, 345, 361, 380, 401-403, 405, 409, 410, 417-421, 426, 427, 437, 449, 461, 462, 493,
 - alloy, 449
 - alloying with copper, 420
 - arsenide (AlAs), 2
 - corrosion, 465
 - silicon diodes, 239
 - spiking, 417-420
- Al_xGa_{1-x}As, 147-148, 159, 292, 304, 307, 310, 334, 357, 381, 385,
- Al_xGa_{1-x}As_ySb_{1-y}, 302
- Al_xGa_{1-x}N, 293, 295
- Ammonia, 405, 406, 425
- Ammonium fluoride(NH₄F), 449
- Amorphous Si (a-Si) solar cell, 345, 346, 347, 353
- Amorphous silicon, 210, 284
- Amorphous surface layer, 488, 497
- Amplifier, 524
- Analog circuits, 512, 524
- Angular frequency, 287
- Anisotropic etching, 451-453, 458, 461, 462, 464
 - profiles, 453, 458, 461, 464
 - reactive sputter etching, 521
- Annealing, 466, 480, 483, 490-495, 499, 501, 504, 522, 532
- Anode, 150, 154
- Antimony (Sb), 411
- Antiphase domain (APD), 387
- Antireflection coating, 327, 328, 330, 332, 337, 343, 352,
- Argon (Ar), 358, 363, 406, 414
- Arc lamps, 493
- Area defect, 372, 373, 388
- Aromatic diamine, 297
- Arsenic, 17, 18, 34, 37, 42, 79, 80, 357, 365-369, 379, 380, 383, 388, 389, 402, 467, 468, 474, 476, 478, 479, 486-488, 491, 495, 496, 498, 499, 501, 503, 504, 512-514, 518, 539
- As (see Arsenic)
- As₂O₃, 467
- AsCl₃, 467
- AsH₃ (see Arsine), 378-380
- a-Si:H TFT, 210, 211
- Aspect ratio, 456,457,460,461,463
- Atmospheric “windows,” 259
- Atmospheric pollution monitoring, 302
- Atmospheric pressure, 377, 381
 - CVD, 400,402,403, 425
- Atmospheric-pollution monitoring, 4
- Atomic layer deposition (ALD), 7, 9, 392, 412, 425, 427
- Attenuation, 300, 308
- Attenuator, 497
- Automobile industry, 1
- Avalanche breakdown voltage, 113-115, 117, 120, 122, 481
- Avalanche multiplication, 82, 111-112, 120
- Avalanche noise, 331, 334
- Avalanche photodiode, 331-333, 352, 355

- B₂H₆, 467
- Backscattering, 444-446
- Ballistic collector transistor (BCT), 149, 150
- Ballistic devices, 277
- Ballistic electron, 274
- Bandgap, 15, 25, 27, 28, 29, 31-33, 38-41, 103-104, 107-108, 110, 112, 115, 117, 119, 122, 142-143, 146-148, 157, 159, 230, 231, 249, 251, 275, 283, 312, 313, 315, 321, 479, 535
- Bandgap narrowing effect, 40, 142

Band-to-band recombination, 56, 78

Barium strontium titanate (BST), 408

Barrel plasma etcher, 456

Barrier height, 230-236, 238, 239, 246, 251, 252, 255, 256, 257 551

Barrier metal layer, 417, 419

Base, 124-131, 133-150, 152, 155, 157-159
region, 126, 130-131, 133, 136-137, 142-143, 148, 158-159
transit time, 140-141, 148
transport factor, 128-129, 143, 146, 157-158
width modulation, 137, 139
width modulation effect, 139

Basic conduction processes, 174

BBr₃, 467

BCl₃, 461, 465

Be, 370, 384

Beam-blanking plates, 441

BF₃, 483

BiCMOS, 195, 205, 210, 226, 524 525

Bidirectional thyristor, 154, 155

Binary compound, 17, 18, 19

Bipolar junction transistor (BJT), 123, 146-147, 159

Bipolar technology, 505, 511, 538,

Bipolar transistor, 2, 3, 4, 5, 7, 10, 12, 82, 111, 123-125, 127-129, 131, 133-135, 137-139, 141, 143-147, 149-151, 153, 155, 157-159, 505, 506, 509, 511-515, 517, 518, 520, 524, 525, 537, 540

Bits, 524, 526

BN, 467

Body-centered cubic (BCC), 19

Boron (B), 189, 190, 194, 223, 359, 361, 369, 375, 381, 389, 390, 402, 467, 468, 473, 476, 477, 478, 480, 482, 486-488, 490-494, 496, 497, 499, 501-504, 512, 513, 515, 518, 522,

Boron implantation, 189

Boron trioxide (B₂O₃), 369

Bow, 372, 376

Bragg reflector, 446
wavelength, 312, 322

Breakdown region, 150

Breakdown voltages, 112-113, 115-116, 120, 262, 279

Bridgman technique, 6, 368, 370

Buffered HF solution (BHF), 449

Buffered-oxide-etch (BOE), 449

Built-in electric field, 143, 148

Built-in potential, 85, 87, 89, 91-94, 99, 119-121, 150, 230, 231, 233, 236, 246, 247, 256, 509

Bulk punch-through, 198, 199

Buried layer, 498, 501, 512-514, 517, 524, 525

Buried-type PMOS, 522

C₂F₆, 461

C₄F₈, 461

C₅F₈, 461

Cadmium, 369

Camera, 529, 536

Capacitance, 476, 495, 500, 508, 510, 512, 514, 518, 527, 535
voltage technique, 476
voltage characteristics, 96, 168

Capture cross section, 61

Carbon, 358, 374, 375

Carrier drift, 43, 77, 79

Carrier generation, 56, 63

Carrier injection, 56, 64

Carrier lifetime, 129, 142, 158, 287, 288

Carrier mobility, 475

Carrier transport, 160, 161, 174, 175

Cathode, 150, 154

Cell projection, 442, 443, 463

Cellular phone, 5, 223, 259, 529

Central processing unit, 8

CF₄, 461

Chain scission, 443

Channel conductance, 181, 184-187, 194, 218

Channel doping profile, 201, 226

Channel length, 180, 195-200, 202 223,

Channel resistance, 184, 201, 202

Channel width, 180, 223

Charge density, 164, 173, 174, 193, 200, 202, 221, 227

Charge distribution, 162, 165, 167, 173, 177, 192, 193, 196

Charge neutrality, 85, 95

Charge sheet, 252, 253

Charge-coupled device (CCD), 5, 160, 177, 178

Charge-sharing model, 196, 197, 226

Charge-trapping devices, 215, 216, 221

Chemical etching, 428, 447, 451, 453, 454, 460, 463, 464

Chemical vapor deposition (CVD), 377, 378, 381, 388, 389, 377, 379, 380, 381, 383, 384, 388, 392, 400, 401, 402, 409, 412, 425, 426-427

Chemical-amplified resist, 437

Chemical-mechanical polishing (CMP), 9, 421-423

CHF₃, 461-462

Chip card, 529, 538

Chromium, 105

Circuit symbol, 125

CIGS, 347, 348, 353, 356

CIS (CuInSe₂), 347, 348

Clean room, 428-430, 437, 459, 464

Clear air turbulence detection, 259

Cleaved ends, 305

Clustered plasma processing, 459

Clustered tools, 459

CMOS (complementary MOS-FET), 6, 7, 8, 9, 14, 193, 195, 196, 199, 202, 203, 205-210, 212, 215, 216, 224-227, 477, 478, 483, 498, 499, 505, 514, 516, 519, 520, 521, 524-526, 532-540
inverter, 205-207, 215, 224, 226, 519, 520, 540
image sensor, 178, 208, 209
technology, 193, 199, 202, 224, 225, 509, 510, 511, 514, 516, 518, 521, 522, 524, 525, 526, 529, 530, 532, 534-540

Cobalt silicide, 423, 424

Cold wall reactor, 401
 Collector, 111, 124-152, 157-159
 -emitter leakage current, 136
 transit time, 140, 149
 Color bands, 280
 Columnar structure, 410, 411
 Common-base configuration, 126, 129, 135
 Common-base current gain, 128-129, 133, 137, 139, 150, 158-159
 Common-base cutoff frequency, 139
 Common-emitter configuration, 135-138, 148, 155
 Common-emitter current gain, 136-137, 140, 145-148, 158-159
 Complementary error function, 471, 472, 497, 501
 Complementary MOSFET (see CMOS)
 Compound semiconductor, 6, 8, 15, 17, 20
 on silicon, 386, 389
 Computer-aided design (CAD), 433, 510
 Concentration-dependent diffusivity, 466, 476, 481, 514
 Conduction band, 25-32, 34-37, 41, 42, 83, 99, 112, 117, 119, 142, 146, 148
 Conduction band discontinuity, 148
 Conductivity, 15, 16, 28, 29, 37, 40, 49, 59, 74, 79, 80, 175, 184
 Conductor, 15, 16, 29, 40
 Confinement factor, 304, 308, 310, 311, 322
 Conformal step coverage, 404, 416, 425
 Constant-mobility regime, 71, 72
 Constant-surface-concentration diffusion, 470, 471, 477, 480, 501,
 Constant-total-dopant diffusion, 470, 473, 481, 484
 Constant-field scaling, 200
 Contact holes, 414, 415
 Contact imaging sensors (CIS), 210
 Contact plug, 417
 Contact printing, 431
 Contact resistance, 482, 501
 Continuity equation, 43, 62, 63, 64, 77, 78, 80, 101, 130, 159
 Contrast ratio, 435
 Conversion efficiency, 264
 Copper (Cu), 19, 105, 449, 461, 462, 524, 528
 halides, 462
 interconnect, 7, 9
 metallization, 421, 425
 Copper indium gallium diselenide (CIGS) (see CIGS)
 Correction factor, 50
 CoSi_2 , 423, 427
 Coulomb force, 45
 Coulombic interaction, 485
 Coupled impurity vacancy pair, 478
 Covalent bonding, 22
 CRAY 1, 10
 Critical angle, 290, 291, 300, 304, 321, 322, 336
 Critical dimension, 431, 463
 Critical dimension (CD) control, 431, 463
 Critical field, 113-115, 263
 Critical layer thickness, 385, 388
 Crystal defect, 372
 Crystal growth, 17
 Crystal growth techniques, 357, 365, 367, 368, 390
 Crystal lattice, 467, 468, 501
 Crystal plane, 15, 21
 Crystal puller, 358, 359
 Crystal purification, 365
 Crystal structure, 15, 17, 20, 40, 41
 Cubic lattice, 372
 Current crowding, 143, 144
 Current density, 43, 48, 52, 54, 55, 65, 80, 81, 235, 236, 237, 256
 equations, 55
 Current gain, 123-124, 127-129, 133-140, 142-148, 150, 152, 155, 158-159, 207
 Current transport, 234, 235, 238, 255
 Current voltage characteristics, 82-83, 99-104, 107-108, 120-121, 129, 131, 135-137, 150-156
 Cutoff frequency, 124, 139-140, 148, 155, 157, 159, 240, 247, 248, 254, 255, 273, 277
 Cutoff mode, 134, 141
 CVD (see chemical vapor deposition)
 for GaAs, 379
 for silicon, 377
 Czochralski crystal, 359, 374, 377
 Czochralski technique, 357, 358, 361, 363, 368, 388, 390
 Damascene process, 462
 Dangling bonds, 68, 210
 Data communication, 259
 De Broglie wavelength, 313, 315
 Deep UV lithography, 437
 Defect, 283, 344, 346, 347, 429, 431, 432, 434, 435, 449, 464
 density, 210, 421, 425, 434, 464, 479, 480, 490, 493, 501, 504, free zone, 377
 Degenerate energy levels, 24
 Degenerate semiconductors, 40, 303
 Degree of anisotropy, 450
 Degrees of freedom, 44
 Density, 361, 364, 377, 383, 386, 388, 389, 390, 392, 394, 395, 397, 403, 405, 407, 408, 414, 419, 420, 421, 422, 425, 426, 427, 434, 441, 449, 451, 454, 456, 457, 461, 463, 464, 465
 Density of states, 30-32, 35, 36, 40, 41, 68, 234, 235, 313, 314
 Denuded zone, 376, 377
 Depletion, 162-166, 168, 177, 179, 181-183, 187, 188, 192, 193, 195-200, 212, 213, 223
 approximation, 164, 166, 199
 capacitance, 82, 95, 96, 98, 110, 120, 121, 139
 case, 162
 layer, 82, 95-97, 109, 115, 120
 layer width, 88-92, 94, 96, 98, 104, 115, 121, 122, 126, 232, 236, 237, 238, 242, 243, 247
 mode, 246, 251, 255, 256
 mode device, 246, 251, 256

region, 82-83, 86-89, 92-93, 95-97, 99-101, 104-106, 109, 112, 113, 120-121, 126, 130, 134, 144, 145, 148, 150
 Depth of focus (DOF), 433, 439, 441, 464
 Developer, 435-437, 443
 Device building blocks, 1, 2
 Device leakage, 196
 DH laser (see double heterostructure laser)
 DI water, 448, 449
 Diac (diode ac switch), 154-156
 Diamond structure, 15, 20
 Diborane, 378
 Dichlorosilane, 377, 403, 404, 406
 Dichlorosilane nitrous oxide, 404
 Dielectric constant, 400, 406-409, 421, 425-427
 Dielectric isolation, 514
 Dielectric layers, 392, 425
 Dielectric permittivity, 232, 251
 Dielectric relaxation time, 267, 268
 Dielectric strengths, 406
 Diethylzinc $Zn(C_2H_5)_2$, 380
 Diffraction, 431, 432, 436, 440, 444, 462
 effect, 436, 440
 Diffusion, 6, 13, 84, 87, 109, 110, 117, 143, 150, 447, 466-485, 492-495, 498, 501-503, 507, 508, 512, 514, 515, 518, 521
 barrier, 406, 417, 421, 424
 capacitance, 96, 109-110, 120, 139
 coefficient, 54, 66, 80, 158, 362, 395, 400, 418, 419, 426, 469, 470, 473, 477, 478, 481, 486, 496, 503
 constant, 130, 141, 158
 current, 53-55, 65, 80, 84, 99, 102, 104-108, 120-122, 137, 143, 145
 equation, 468, 469, 471, 474, 475, 477, 480, 484, 487, 497, 501, 511,
 in silicon, 467-469, 473, 478, 479, 480, 482, 486, 487, 488, 490, 499, 501-503
 length, 65, 80, 102, 109, 121, 122, 124, 130, 131, 155, 471, 473, 474, 479, 503, 505, 507, 516, 517, 518, 523, 525, 532, 534, 535, 537, 539, 540
 length of electrons, 101
 length of hole, 101, 130
 mask, 117
 process, 53, 117, 468, 469, 471, 474, 475, 477, 480, 484, 487, 490, 497, 501, 511
 profiles, 470, 472, 475, 476-480
 in strained Si, 479, 502
 Diffusivity, 54, 55, 80
 Digital cameras, 5, 208, 529
 Digital circuits, 512
 Digital video disk, 4
 Dipole layer, 269
 Direct exposure, 441
 Direct recombination, 56, 60, 62
 Dislocation, 372, 373, 374, 376, 377, 385-389, 391, 429
 Dissociation effect, 478, 501
 Dissolution of silicon, 419
 Distributed Bragg reflector laser, 312, 313
 Distributed feedback (DFB), 319
 laser, 312, 313
 D-MOSFET, 223
 Donor, 15, 33-39, 41, 42
 Donor concentration, 234, 241, 256
 Donor states, 170, 171
 Donor vacancy, 476
 Dopant profile, 473, 474, 476, 498, 501
 Doping distribution, 361, 363, 364, 389
 Double barrier structure, 270, 271
 Double charged acceptor vacancy, 476
 Double-drift IMPATT diode, 262
 Double heterojunction, 286, 287, 303, 319
 Double-heterostructure (DH) laser, 303, 304, 307, 308, 310, 311, 312
 Drain, 180-188, 190-192, 194, 195-205, 208, 210-214, 218, 219, 223, 227
 -induced barrier lowering (DIBL), 197-199
 DRAM (see dynamic random access memory)
 Drift current, 48, 53, 55, 71, 80, 181
 Drift region, 223, 262-264, 279
 Drift velocity, 44, 45, 51, 55, 66, 73-75, 78, 81, 529, 532
 Dry etching, 7, 8, 428, 437, 450, 451, 452, 454, 460, 462, 463, 465
 Dry oxygen, 393, 395, 396, 397, 399
 Dry photoresist stripping, 439
 DSSC (dye-sensitized solar cells), 348, 349
 Dual damascene process, 421
 Dust particle, 429, 430, 431, 464
 Dynamic random access memory (DRAM), 7, 8, 10, 12, 408, 507, 526, 527
 Earth's crust, 16
 Early effect, 138
 Early voltage, 137-138
 E-beam evaporation, 414, 427
 Edge dislocation, 372, 385, 388
 Effective density of states, 32, 35, 36, 40, 41
 Effective mass, 26-28, 31, 34, 41, 42, 44-46, 69, 71, 74, 76, 236, 238, 239
 Effective segregation coefficient, 362, 364
 Effusion oven, 381, 383, 390
 Einstein relation, 54, 55, 336
 Elastic collision, 485
 Elastic forces, 385
 Electrically erasable programmable read-only memory (EEPROM), 215, 219
 Electrochemical anodization, 392
 Electrochemical methods, 421
 Electroluminescence phenomenon, 3
 Electromagnetic spectrum, 280, 281
 Electromigration, 9, 417, 420, 421, 422, 427, 462
 Electron affinity, 68, 69, 117, 119, 160, 189, 195, 229, 230, 256

- Electron bombardment, 172
- Electron cyclotron resonance (ECR), 457,458
- Electron gun, 441
- Electron mobility, 45, 46, 74, 79, 80
- Electron resist, 433, 443
- Electron temperature, 274, 276, 279
- Electron-beam lithographic system, 433
- Electron-hole pair generation, 104
- Electronic industry, 1, 2, 12
- Electronic purse, 529
- Electronic stopping, 485-487
- Electronic-grade silicon, 358
- Element semiconductor, 16, 17, 20
- Emission lines, 312
- Emission processes, 104
- Emission wavelengths, 292, 302, 313
- Emitter, 124-150, 155-159
 - efficiency, 128-129, 133-134, 145, 148, 157-158
 - bandgap narrowing, 142
- End-point control, 459, 460
- Energy barrier thickness, 270
- Energy momentum diagram, 26, 27
- Enhancement-mode MOSFET, 519
- Epitaxial film, 429
- Epitaxial growth, 6, 8, 378, 379, 385, 388,
 - technique, 357, 377, 390
- Epitaxial layer, 240, 246, 256,512, 513, 515, 524, 532
- Epitaxial process, 377, 380, 384, 388
- Epitaxy, 357, 377, 379, 380, 382, 384, 385, 386, 388, 389
- Equilibrium segregation coefficient, 359, 361, 369, 370
- Equivalent circuit, 110, 138, 139, 510
- Erfc (see error function)
- Error function, 471, 472, 497, 501
- Etch anisotropy, 461
- Etch chemistry, 455, 460
- Etch mechanism, 454
- Etch rate,403, 406, 447-451, 453-455, 457, 460-465
- Etch selectivity, 453-454, 460, 461, 463, 465
- Etch solution, 447,449
- Etching, 407, 416, 421, 422, 428, 437, 438, 447-465
 - equipment, 460
 - operations, 514
- Evaporation, 379, 381, 383, 388, 414, 427
- Exclusive NOR, 273, 274
- Exhaust system, 358
- Exposure response curve, 435, 436
- Exposure tools, 428
- Extreme-ultraviolet (EUV) lithography, 445, 446
- Extrinsic ,
 - base regions, 503
 - diffusion, 476, 478, 479, 503
 - semiconductor, 49
 - stacking fault, 373, 374
 - transition, 283
- Eye response, 291
- External quantum efficiency, 289, 319
- F-doped oxide, 535
- Fermi distribution, 30, 31, 36
- Fermi level, 15, 30, 31, 33, 35-39, 42
- Fick's diffusion equation, 6, 469
- Field oxide, 392, 398, 402, 425, 426
- Field transistor, 190, 194
- Field-effect transistor (FETs), 377, 384, 388
- Film formation, 505, 514
- Film thickness, 436, 437, 451, 460
- FinFET technology, 525
- Fixed charge, 172, 189, 227
- Fixed-oxide charge, 170, 172-174, 181, 188
- Flash memory, 215, 219, 220
- Flat doping profile, 495
- Flat-band condition, 161, 163, 169, 170, 173, 193
- Flat-band voltage, 170, 172-174, 188, 193, 194
- Flat-panel display, 291, 295, 319
- Flip-flop structure, 215
- Floating dot, 219
- Floating gate, 5, 215-219, 221, 226, 227, 528, 540
- Float-zone process, 357, 363-365, 388, 390
- Float-zone silicon, 366
- Flow control, 358
- Focused electron beam, 414
- Forward breakover, 150, 153
- Forward-blocking state, 153
- Four-point probe method, 50, 374
- Free-carrier flux density, 119
- Frenkel defect, 372
- Frenkel-Poole emission, 175
- Frequency locking circuit, 260
- Frequency response, 123, 137, 138, 141, 143, 149, 155, 159, 287, 288
- Fresnel reflection loss, 290
- Full width at half maximum (FWHM), 271, 287
- Full-color displays, 293
- Full-color indicators, 293
- Fused quartz, 15
- Fused silica, 433, 434
- Fused-silica crucible, 369
- Ga, 449, 467
- $Ga_{0.47}In_{0.53}As$, 248, 254
- $Ga_{0.51}In_{0.49}P$, 344
- $Ga_{1-x}In_xAs$, 385
- GaAs (see gallium arsenide)
- $GaAs_{1-y}P_y$, 18
- Gain factor, 309, 310, 322
- GaInAs/AlInAs material, 275
- GaInAs-AlGaAs DH, 312
- $Ga_xIn_{1-x}N$, 291, 292, 295
- GaInP/GaAs tandem cell, 353
- Gallium arsenide (GaAs), 2, 6, 17, 18, 22, 27, 28, 32-35, 39, 40, 41, 44, 46, 50, 55, 56, 74, 75, 77, 231, 233, 236, 240, 246, 255, 260, 261, 265, 268-271, 273, 275, 278, 279, 285, 288, 289, 291, 292, 303-305, 307, 308, 310, 311-314, 318, 320-322, 334-337, 343-345, 353, 354, 357, 367-371, 377, 379, 381, 383, 384, 387, 389, 390, 392, 393, 467, 469, 470, 476, 479, 486, 488, 496, 501, 502, 529, 530, 532
- Gallium arsenide etching, 449
- Gallium chloride, 379

Gallium phosphide (GaP), 289-292
 Gallium-arsenic system, 367, 368
 GaN, 293, 295
 GaP, 17, 18
 Gas source, 358
 Gas thyatron, 123
 GaSb, 260, 261, 278
 Gate, 153, 155, 163, 169, 176-183,
 186-197, 199, 200, 219, 221-
 223, 225-227
 capacitance, 183, 202, 221, 227
 current, 153, 155
 electrode, 150, 153
 length, 240, 248, 253-255 oxide,
 518, 519, 526-528, 532, 533,
 536, 539
 oxide thickness, 518, 519, 526,
 527, 528, 532, 533, 536, 539
 stack, 201
 Gaussian distribution, 473, 475,
 484, 488, 495, 501, 514
 Gaussian spot beam, 442
 $Ga_xIn_{1-x}As_yP_{1-y}$, 298, 302
 $Ga_xIn_{1-x}As_ySb_{1-y}$, 302
 Ge_{0.3}Si_{0.7} on silicon, 385
 Generation lifetime, 104
 Generation-recombination centers,
 104-105, 118, 121, 142
 Generation-recombination
 rates, 168
 Germanium (Ge), 260, 261, 277,
 285, 300
 Gettering, 376
 Ge_xSi_{1-x} , 385
 G-line, 433
 Global planarization, 392, 421, 425
 Gold, 19, 105, 110, 446, 478
 contact, 246
 doping, 207
 silicon photodetector, 330
 Graded base, 141, 143, 148-149
 Graded layer, 148-149
 Graded profile, 148
 Graded-index fiber, 300
 Gradual-channel approximation,
 195
 Grain boundary, 210, 211, 373
 Graphite boat, 368
 Graphite crucible, 369, 377
 Graphite susceptor, 358, 374, 380
 Greek alphabet, 546
 GRIN-SCH structure, 315
 Groove, 371
 Gross world product (GWP), 1, 2
 Gunn diode, 4

 H_3PO_4 , 449, 450
 Half life, 366
 Hall coefficient, 52, 79
 Hall effect, 50, 51, 52
 Hall field, 51, 52
 Hall measurement, 50, 52, 79
 Hall voltage, 51, 52, 79
 Haynes-Shockley experiment,
 66, 67
 HDP, 451, 457, 461
 Heater, 153
 Heavy holes, 313
 Helium, 302
 -neon gas laser, 302
 -neon laser, 460
 Heteroepitaxial growth, 380, 387,
 388
 Heteroepitaxy, 357, 384, 386, 388
 Heterojunction, 82, 117-122, 127,
 147, 148, 297, 301, 304,
 bipolar transistor, 3, 4, 123, 146-
 147, 149, 155
 field-effect transistor, 384
 interface, 1, 2
 Heterostructure,
 laser, 3, 4
 Hexa-methylene-di-siloxane
 (HMDS), 437
 HF, 447-450, 464, 474, 503
 HFET, 249, 257
 HfO_2 , 409, 414, 417
 High current implantation, 466,
 495, 498, 501
 High dielectric-constant materials,
 214
 High electron mobility transistor
 (HEMT), 249, 256
 High energy implantation, 466, 495,
 498, 501
 High injection, 104, 106, 107, 120,
 142, 143, 146
 High-aspect feature, 414

 High-current densities, 107
 High-density plasma (HDP), 457,
 461, 463, 465
 etching, 465
 High-fidelity transfer, 451
 High-field effects, 43, 73
 High-frequency equivalent circuit,
 138-139
 High-*k* dielectric materials, 534
 High-*k* materials, 408, 409
 High-resolution gas spectroscopy,
 302
 High-speed laser printing, 302
 High-voltage direct current
 (HVDC) devices, 157
 Hi-lo structure, 262
 H-line, 433
 HNO_3 , 474
 Hole, 23-25, 27-29, 31, 32, 34-38,
 40, 41, 42
 Hole ionization rates, 113
 Hole mobility, 45
 Homoepitaxial growth, 384
 Homoepitaxy, 357, 384, 388
 Homojunction laser, 303
 Hot carrier injection, 215, 219
 Hot electron device, 258, 274, 277,
 278
 Hydrofluoric acid, 447-448
 Hydrogen (H_2), 358, 377, 378, 380,
 401, 406, 417, 426, 484, 488,
 500
 annealing, 170
 atom, 23, 34
 chloride (HCl), 358
 Hydrogenation treatment, 211
 Hyperabrupt junction, 98

 IC card, 529
 IC design, 434
 IC processing, 434, 439, 447, 462
 Ideal characteristics, 99, 120
 Ideal diode current, 342
 Ideal diode equation, 102-107
 Ideal MOS curves, 165
 Ideality factor, 106
 IGFET, 180, 225
 III-V compound semiconductor,
 240, 255

III-V nitride LED, 293, 295
 I-line, 433
 Impact ionization, 43, 76-78
 Immersion lithography, 7, 9, 14, 441
 IMPATT diode, 258, 260-265, 267, 277-279
 Implant damage, 466, 490, 492, 494, 499, 501, 504
 Impurity distribution, 475, 481
 Impurity redistribution, 466, 481, 482, 501
 Impurity scattering, 45
 Incandescent lamps, 298
 Index of refraction, 433
 Indirect recombination, 56, 60, 62
 Indium phosphide (InP), 18, 228, 248, 254, 265, 268, 274, 279
 Inductively coupled plasma etcher, 456, 457
 Infrared LED, 286, 291, 298, 299, 319, 321
 Infrared sensors, 323
 Infrared system, 259
 InGaAsP-InP, 301
 InP (see indium phosphide)
 InP/InGaAs structure, 148
 Input conductance, 139
 Insulator, 15, 16, 28, 29
 Integrated circuit, 2, 4, 6, 7, 9, 12, 13, 83, 155, 160, 189, 193, 195, 200, 205, 223, 224, 357, 358, 373, 376, 377, 388, 392, 398, 400, 414, 416, 417, 420, 423, 425, 428, 434, 463, 474, 497, 501, 505-509, 532, 538
 technology, 392
 inductor, 510, 511
 interacting *p-n* junctions, 126
 Interconnect, 459, 461-463
 Interconnections, 392, 406, 423, 426
 Interface, 481, 482, 493, 500, 501, 504
 trap density, 170, 172, 221, 425
 -trapped charges, 170, 173
 trap levels, 170, 171
 trap system, 170, 171
 Interlayer dielectric (ILD), 421
 Internal quantum efficiency, 288, 289, 321
 Interstitial diffusion, 468, 469
 Interstitial site, 372, 390
 Interstitials, 468, 480, 493
 Intrinsic,
 base region, 515
 carrier density, 32, 33, 40, 108
 diffusion, 476, 478
 diffusivity, 476, 478, 501
 Fermi level, 33, 36, 37, 42, 47, 59, 104, 121, 160
 semiconductor, 29-33
 stacking fault, 373, 374
 transitions, 283, 285
 vacancy density, 477
 Inversion, 160, 162-168, 180-185, 187, 188, 191-193, 199-200, 202, 203, 211, 223,
 case, 162, 163
 layer, 163, 164, 168, 181, 183-185, 187, 199, 200, 202, 203, 211
 staggered structure, 210
 Inverted mode, 134
 Ion beam, 476, 484, 488, 489, 497, 498, 503
 Ion channeling, 488
 Ion distribution, 466, 483, 484, 498, 502
 Ion implantation, 6, 83, 87, 143, 437, 466, 467, 483, 485, 493, 494, 498, 499, 501, 502, 503
 Ion stopping, 485
 Ion-beam lithography, 446, 462
 Ionic conduction, 175
 Ionization energy, 34, 76
 Ionization integrand, 261
 Ionization rate, 77
 Isoelectronic, 288, 293, 294
 Isopropyl alcohol, 448
 ITO (indium tin oxide), 297, 298
 Junction breakdown, 82, 83, 111, 112, 120, 122
 Junction curvature effect, 117
 Junction depth, 180, 196, 201-204, 474, 475, 477, 498, 501, 502, 503
 Junction spiking, 418
 Kinetic energy, 29, 41, 44, 48, 69, 76, 78
 Kink, 213, 478
 Kirchhoff's circuit laws, 125
 KOH, 448, 450, 464
 KrF excimer laser, 433
 Lamps, 291, 293, 295, 298
 Lanthanum hexa-boride (LaB₆), 441
 Lapping, 447
 Large-scale integration, 507, 532
 Laser crystallization of Si, 211
 Laser interferometry, 460
 Laser printing, 4
 Lasing modes, 306
 Latch-up, 203, 206-208, 212, 225, 520, 521
 Lateral diffusion, 466, 380, 481, 501, 518, 521
 Lateral oxide isolation, 512, 517
 Lateral straggle, 484, 487
 Lattice, 17-22, 26, 27, 34, 41, 84, 112, 118
 constant, 19-21, 41, 119
 scattering, 45, 46, 79
 matched, 357, 384, 385, 388
 mismatch, 385-387
 Lead zirconium titanate (PZT), 408, 409
 Leakage current, 181, 194, 198, 199, 203, 205, 207, 214, 227, 519, 521, 522, 526, 534
 Lifetime, 43, 57-59, 67, 80, 109, 142, 157, 158, 207, 491
 of minority carriers, 109, 110, 130
 Liftoff technique, 437
 Light bulb, 153
 Light holes, 313
 Light-emitting diodes (LEDs), 280, 286-293, 295, 298, 299, 301, 306, 312, 319-321
 Lightly doped drain (LDD), 498
 Limited-source diffusion, 514
 Line defect, 372, 373, 388
 Linear rate constant, 397-399
 Linear region, 183-185, 196, 226, 227, 244, 245, 253, 255

Linearly graded junction, 87-88, 92-94, 97-98, 114-115, 121
 Liquid crystal displays (LCD), 210
 Liquid encapsulant, 369, 388
 Liquid encapsulation method, 368
 Liquidus line, 367, 368
 Lithographic operations, 514
 Lithographic step, 229
 Lithography, 83, 428, 429, 431-433, 437, 439-442, 444-447, 449, 462-464
 Lithography process, 6, 7
 Load devices, 215, 225
 Load line, 138, 141
 Load transistors, 215, 216
 Local oscillator, 260, 265, 277
 LOCOS process, 514
 Logic circuits, 195, 205
 Longitudinal field, 248, 253
 Lorentz force, 51
 Loss, 289, 290, 306, 308, 312, 320, 321
 Low dielectric constant, 400, 406, 407, 421, 425
 Low pressure, CVD (LPCVD), 377, 400, 416 reactor, 410
 Low-dielectric-constant (low-k) materials, 400, 406, 407, 408, 421, 510, 528, 535
 Low-*k* materials, 407, 408
 Low-resistivity metals, 510
 Magnetically enhanced RIE (MERIE), 451, 455
 Majority carrier concentration, 43, 62, 79
 Majority carrier current, 236, 238
 Mask, 428-429, 431-442, 445-446, 448-451, 460, 462-465
 beam system, 446
 blank, 445, 446, 462
 thickness, 496
 Masking, 466, 495, 496, 497, 501, 504
 Maskless direct writing, 441
 Mass action law, 100
 Maximum output power, 340
 Maximum power, 339-342, 356
 Maximum time to breakdown, 409, 410
 MBE (see molecular beam epitaxy), 377, 380, 381
 Mean free path, 44, 46, 53, 382, 383, 390
 Mean free time, 44-46, 53
 Mean time to failure (MTF), 420
 Medium-scale integration (MSI), 507
 Melting point, 368, 369, 377, 388
 Memory devices, 505, 526
 Mercury-arc lamp, 433
 MERIE, 451, 455, 461
 MESFET, 228, 240, 241-249, 254-256
 integrated circuit, 532
 technology, 529, 540
 Metal, 22, 28, 29, 69, 81
 film, 392, 405, 414, 425
 interconnect, 459, 461, 463
 semiconductor contact, 228-231, 233, 234, 236, 238, 255
 semiconductor device, 261
 Metallization, 392, 402, 406, 414, 416, 417, 419, 421, 423, 425, 426, 427, 505, 513, 517, 519, 524, 540
 Metallurgical-grade silicon, 358
 Metal-nitride-oxide-semiconductor (MOS), 216, 221-223, 529
 Metalorganic chemical vapor deposition (MOCVD), 7, 8, 381, 389
 Metalorganic compound, 380, 384, 388
 Metalorganic molecular-beam epitaxy (MOMBE), 384
 Metal-oxide-semiconductor (MOS) structure, 2
 Metal-semiconductor contact, 3, 12
 Metal-semiconductor interface, 1, 2
 Metal-semiconductor photodiode, 327, 330
 Metal-SiO₂-Si, 165, 168, 169, 193
 Metal work function, 229, 230, 256
 Metric system, 429
 Microcrystalline/amorphous tandem cell, 347
 Microelectromechanical systems, 428
 Microelectronics, 464
 Microprocessor, 7, 8, 10, 11, 12
 Microprocessor chip, 507, 524
 Microprocessor-based control system, 358
 Microwave region, 258
 Microwave technology, 259
 Midgap, 163
 Miller indices, 21, 22, 40, 41
 Millimeter wave band, 259, 277
 Millimeter wave devices, 258
 Millimeter-wave power amplifiers, 146
 Minimum feature length, 9, 517, 532, 535
 Minimum linewidth, 431, 445
 Millisecond annealing, 494, 495
 Minority carrier, 43, 55-60, 63, 65-67, 78, 80, 163, 168, 177, 203, 207, 223
 Minority carrier current, 236-238
 Minority-carrier lifetime, 374
 Minority-carrier storage, 109-110, 120
 MIP (million instructions per second), 10
 MISFET, 180
 Misorientation, 488, 490
 Mix-and-match approach, 446, 462
 Mixers, 146
 MNOS transistor, 221
 Mobile ionic charge, 170, 172, 173
 Mobile radio, 510
 Mobility, 43, 45, 46, 50, 55, 66, 71-74, 77-81, 148, 149, 157, 181, 184, 201-212, 225, 475, 479, 491, 501, 502, 507, 512, 520, 529
 MOCVD, 380, 381, 383, 384, 388
 MODFET, 228, 249-255, 257
 Modulation bandwidth, 288, 312
 Modulation frequency, 312
 Modulation-doped field-effect transistor (MODFET), 3, 5
 Mole fraction, 378
 Molecular beam epitaxy (MBE), 7, 8, 270, 278, 279, 377, 380,

382, 384, 389
 Molecular impingement rate, 381
 Molecular weight, 381, 383, 390
 Molecular-beam epitaxy, 377, 380, 382, 384, 388
 Monolithic IC, 7
 Monolithic microwave integrated circuits (MM IC), 509, 532
 MOS capacitors, 508
 MOS devices, 453, 461, 512, 535, 536, 538
 MOS memory cells, 443
 MOSFET, 2, 3, 4, 5, 8, 12, 195-227, 388, 392, 393, 397, 420, 423, 425, 505, 506, 516-520, 525, 527, 528, 532, 534, 538-540
 MOSFET scaling, 195, 200, 201, 209, 213, 214, 224, 226,
 MOSFET technology, 505, 516, 539
 MOSi₂, 424
 MOST, 180
 Multilevel metallization, 392, 402
 Multiple implantation, 495, 501
 Multiple-quantum-well (MQW), 314, 315, 318

 NbN, 414
n-channel MOSFET, 180, 187, 189, 190, 195-198, 200, 203, 205-207, 213, 215, 218, 226
 Near-infrared region, 327
 Negative differential mobility, 74, 265, 268
 Negative photoresist, 435-437, 443, 464
 Negative resistance, 4, 260, 263, 265, 278
 Neutron, 366, 367, 388, 390
 Neutral beam plasma etcher, 457, 458
 Neutron irradiation, 207, 366, 367, 388
 NH₃, 405, 417
 NiSi, 423, 534
 Nitric acid, 447-449
 Nitrided oxide, 522
 Nitrogen, 293, 294, 297, 319
 Nitrous oxide, 403, 404
 Noise factor, 331, 332

 Nonconformal step coverage, 404
 Nonequilibrium situation, 56
 Nondegenerate semiconductor, 35
 Nonideal effects, 123, 142
 Nonpolar semiconductor, 387
 Nonvolatile semiconductor memory (NVSM), 5, 8, 10, 226, 227
 Notebook computer, 5
n-tub, 520, 521, 540
 Nuclear collisions, 488, 490, 491
 Nuclear fusion reaction, 336
 Nuclear spectroscopy, 259
 Nuclear stopping, 485-487, 490
 Nuclear stopping power, 485, 486
 Numeric method, 94
 Numerical aperture, 432, 433, 441
n-well, 190, 206

 Offset voltage, 148
 Ohmic contact, 1, 2, 4, 228, 229, 238, 239, 240, 255, 293, 323, 337, 392, 410, 425
 On condition, 141
 One-sided abrupt junction, 90-91, 96, 112, 114-116, 122
 Open tube, 467, 468, 492
 Optical absorption, 283-284, 328, 345
 Optical concentration, 323, 352
 Optical diffraction, 431
 Optical emission spectroscopy, 459
 Optical fiber, 299-300, 301, 316
 Optical fiber communication, 4, 286, 291, 301, 302, 312, 319, 320, 323
 systems, 307, 352
 Optical gain, 305, 306, 315, 316
 Optical lithography, 9
 Optical lithographic system, 437, 464
 Optical proximity correction (OPC), 440
 Optical reading, 302, 319
 Optical resonant cavity, 283
 Opto-isolator, 291, 298, 334, 352
 Organic LED (OLED), 295-298
 structure, 297
 Organic semiconductors, 295-297

 Organic solar cells, 349
 Organic solvent, 435, 438
 Orientation dependent, 397
 etching, 448, 465
 Oscillator, 524, 535
 Oxidation, 83
 Oxide layer, 229
 Oxide masking, 6
 Oxide thickness, 176, 177, 180, 190, 192, 194, 201, 218, 221, 227
 Oxide trapped charge, 170, 172-174, 193
 Oxygen, 374-377, 381, 382
 Oxygen molecule, 396
 Ozone (O₃), 403, 425

p⁺-polysilicon, 169, 189
 Pb, 18, 29
 PC, 10
 P₂O₅, 467
 P-glass flow, 404, 405, 427
 Percolation theory, 176
 Periodic table, 16, 17, 19
 Phase diagram, 367, 368
 of the Al-Si system, 418
 Phase-shifting mask (FSM), 440
 Phosphine (PH₃), 378
 Phosphoric acid, 449
 Phosphorus (P), 359, 366, 375, 388, 402-405, 410, 411, 425, 467, 468, 478, 479, 480, 482, 486, 487, 492, 496, 501, 503, 517-519, 525, 526
 doped silicon dioxide, 402, 404, 425
 doped oxide (P-glass), 519
 Photo-acid generator, 437
 Photoconductivity method, 374
 Photoconductor, 323-325, 352, 354, 355
 Photocurrent gain, 325, 333, 334
 Photodetector, 281, 323, 328, 330, 335-336
 Photodiode, 298, 325, 327-334, 352, 353, 355, 356
 Photomask, 429, 431, 434, 436, 437, 441
 Photon energy, 280, 283, 285, 291
 Phototransistor, 333, 334, 352

Photonic application, 377, 386
 Photonic devices, 2, 8
 Photoresist, 6, 12, 13, 435-439, 443, 449, 454, 462, 464, 465, 512, 513, 518, 532, 540
 Photoresist adhesion, 449
 Photosensitive compound, 435
 Physical etching, 456
 Physical sputtering, 453, 455
 Physical vapor deposition, 414
p-i-n photodiode, 328-330, 355
 Pinch-off voltage, 244, 246, 251, 257
 Pinch-off point, 183, 184, 187
 Planarization, 392, 403, 407, 421, 422, 425, 426
 Planar process, 6, 7, 83, 153
 Plasma damage, 421
 Plasma etching, 451-453, 457, 462, 464, 465
 Plasma-enhanced chemical vapor deposition (PECVD), 400
 Plasma nitride, 406
 Plasma reaction, 393
 Plasma reactor, 211
 Plasma spray deposition, 414
 Plasma-assisted etching, 428, 451
 Platinum silicide (PtSi), 239
 Platinum (Pt), 19, 414
 PLZT, 409
 PMN, 409
p-n junction, 2, 3, 4, 7, 12, 47, 56, 76, 78, 82-87, 91-92, 95, 98-99, 101, 103, 113, 117, 120-126, 139, 145, 150, 153, 155, 157, 160, 164, 166, 178, 181, 191, 203, 208, 261, 273, 286, 303, 291, 325, 326, 344, 348, 476, 495, 502, 508, 509, 511
 solar cell, 337, 338, 356
 POCl₃, 467
 Point defect, 372, 373, 388
 Point-contact transistor, 3
 Poisson's equation, 63, 71, 85, 87, 88, 90, 92, 95, 113, 164, 172, 266, 267
 Polar semiconductor, 387
 Polishing, 447
 Polonium, 19
 Poly-butene-1 sulfone (PBS), 443
 Polycide, 423, 424, 518
 Polycrystalline GaAs, 388
 Polycrystalline rod, 363
 Polycrystalline silicon, 169, 192, 210, 358, 363, 387, 392, 425
 Poly-glycidyl methacrylate-co-ethyl acrylate (COP), 443
 Polyimide, 535, 540
 Polymer linking, 435, 443
 Polymer deposition, 453
 Polymer LED (PLED), 296
 Polymers, 296, 433, 443, 455
 Poly-methyl methacrylate (PMMA), 443
 Polysilicon, 169, 173, 179, 180, 186, 192, 193, 194, 201, 202, 210, 211, 212, 215, 218, 219, 226, 227, 392, 400, 403, 404, 409-411, 416, 420, 423, 424, 425, 427, 447, 449, 460, 461, 465
 Polysilicon TFT, 210, 211, 212, 215
 Population inversion, 283, 303, 305, 313, 318
 Positive resist, 435-437
 Power amplifier, 265
 Power consumption, 160, 200, 205, 214, 224,
 Power supplies, 155
 Precipitate, 374, 376
 Predeposition, 474, 498, 501, 502,
 Pressure sensor, 454
 Primary flat, 370, 372
 Principal quantum number, 24
 Probability of occupying, 30
 Projected range, 485, 487, 488, 491, 498, 499, 501
 Projected straggle, 484, 487, 501
 Projection printing, 431, 432, 464
 Proximity effect, 440, 444, 462
 Proximity printing, 431, 464
p-tub, 520, 540
 Pt film, 515
 Pyrometer, 493
 Pull rate, 358, 363
 Punch through, 115-116, 148, 521, 535
 diode, 115
 PVD, 404, 414, 416, 417, 421, 425
p-well, 206, 207
 Pyramids, 343
 PZT, 408, 409
 Quality factor, 510
 Quantum-cascade laser, 317, 318
 Quantum dot
 laser, 315-317, 320
 solar cells, 351, 354
 Quantum-effect devices, 269, 273
 Quantum efficiency, 287- 289, 293, 307, 314, 319, 322
 Quantum tunneling, 258, 260
 Quantum well infrared photo-detector (QWIP), 335, 336
 Quantum well thickness, 270
 Quantum-well lasers, 313, 314
 Quantum wire laser, 316
 Quartz, 210, 211, 219, 467, 493
 windows, 493
 Quartzite, 358
 Quasi-Fermi level, 59, 60
 Quaternary compound, 17, 18
 Quaternary, 291, 298, 302, 334
 QW lasers, 313
 Radar system, 259, 260
 Radiation-damage toleration, 212
 Radiative transitions, 280, 281, 293, 321
 Radio astronomy, 259
 Radio frequency, 451, 456, 457, 509, 538
 Radius of curvature, 481
 Raised source/drain, 204
 Random access memory, 7, 8, 10, 507, 526, 527
 Range, 74, 79, 467, 469, 480, 483, 484, 485, 487, 488, 490, 491, 492, 494, 498-504
 Rapid thermal annealing (RTA), 492-494, 501
 Rapid thermal processing, 494
 Raster scan, 442, 443
 RC time constant, 406, 422
 RC time delay, 535
 Reactive ion etching (RIE), 421
 Reactive ion etcher, 454-456
 Reactor, 358, 377-381, 393, 400-402, 406

Real space transfer (RST), 275
 Recombination, 43, 56, 57, 60-65, 68, 77, 78, 80
 center, 56, 60-62, 68, 80
 current, 291
 rate, 56, 57, 60-63, 288,
 Recrystallization, 492
 Rectification, 82
 Rectifying contact, 1, 2
 Rectifying metal semiconductor barrier, 392
 Redistribution diffusion, 474
 Redistribution process, 481, 482
 Reflection loss, 290
 Reflectivity, 305, 316, 321, 322, 347, 355
 Refractive index, 291, 300, 303-305, 312, 328, 343, 440, 441, 460
 Registration, 431
 Reliability, 409, 426
 Resin, 435, 437
 Resist, 428, 431, 433, 435-439, 441-446, 449-451, 456
 Resistance, 228, 229, 238-242, 247, 507-510, 512, 518, 525, 527, 528, 529, 532, 534, 536, 539
 heating, 414
 Resistivity, 29, 43, 47, 49-51, 79, 364, 366, 367, 374, 375, 377, 388, 389, 462
 of silicon nitride, 406
 Resolution, 428, 431-433, 435, 437, 439, 440, 441, 443, 444, 446, 451, 462, 463, 464
 enhancement techniques, 428, 433, 439, 462
 Resonant circuit, 99
 Resonant frequency, 99
 Resonant tunneling diode (RTD), 3, 5, 258, 269, 270, 272, 277, 278
 Response speed, 327
 Responsivity, 327-329, 354, 355
 Retrograde channel profile, 201
 Retrograde well, 207, 521
 Reverse substrate source bias, 188, 191
 Reverse-blocking state, 150, 153
 Reverse-breakdown region, 150
 364, 369, 370, 387
 Selectivity, 453, 455, 456-463, 465
 Selenium, 369, 380, 467
 Self-aligned metal silicide, 423
 Self-aligned gate process, 8
 Self-aligned silicide source/drain contact, 201
 Self-aligned structure, 514, 515
 Semiconductor controlled rectifier (SCR), 150
 Semiconductor industry, 1, 6, 8, 12
 Semiconductor lasers, 280, 283, 302, 303, 307, 320, 321
 Semiconductor material, 15-18, 40, 47, 53, 62
 Semiconductor power devices, 120
 Semiconductor surface, 160-163, 166, 170, 173, 183, 190, 197, 198, 201, 204, 221, 226
 Semiconductor technologies, 1, 6,
 Semiconductor work function, 229, 235
 Semiinsulating GaAs, 369
 Semiinsulating substrate, 242, 246
 Sensor, 433
 Separation by implantation of oxygen (SIMOX), 499
 Series resistance, 106, 120, 238, 247, 287, 291, 330, 342, 343, 345, 350, 509, 512, 525, 536
 Shadow effect, 498
 Shadow printing, 431
 Shallow trench, 540
 Shallow trench isolation, 460
 Sheath, 452, 454, 455
 Sheet charge, 172, 173
 Sheet resistance, 474, 475, 501, 540
 Short-channel effects, 195, 196, 200, 201, 202, 203, 204, 224
 Si, 358, 359, 361, 366, 369, 370, 374, 377, 384-388, 391,
 Si/SiGe HBTs, 148
 Si/SiGe material, 148
 Si₃N₄, 509, 518, 525, 526
 SiC, 292, 293
 Sidewall passivation, 453, 461-462
 SiGe MODFET, 254, 255
 Silane, 377, 380, 402-406, 410, 411, 417, 425-427
 Reverse-breakdown voltage, 153
 RF heating, 401, 414
 RF sputtering, 414, 416, 452
 Richardson constant, 236
 RIE, 451, 453, 454, 461
 Ring oscillator, 535
 Rotation mechanism, 358
 RST transistor (RSTT), 275
 Ru, 414
 Ruby laser, 302
 Salicide, 423, 424, 427
 Sand, 358, 387
 Sapphire substrate, 293, 295
 Satellite, 336, 337, 343, 353
 Saturation channel current, 248
 Saturation current, 102-103, 108, 121, 126, 130, 159, 186, 203, 236, 242, 253, 256
 Saturation current density, 102, 108, 236, 256
 Saturation mode, 134-136, 141-142, 153
 Saturation region, 183, 185, 186, 197, 244-246, 253
 Saturation velocity, 55, 73, 78, 263, 264
 Saturation voltage, 242, 245, 253
 Scaling of MOSFET, 200
 Scaling rules, 200
 Scanning electron microscope, 404, 405
 Scanning focused-beam system, 446
 Schottky barrier, 229, 234, 236, 240, 241, 251, 252, 255, 256, 257
 diodes, 231, 233, 234, 236, 256
 emission, 175
 source/drain, 203
 Sealed ampules, 467
 Secondary flat, 371, 372
 Secondary ion mass spectroscopy (SIMS), 374, 476
 Seed, 358, 363, 364, 369, 370, 377, 387-390
 crystal, 357-365, 367-374, 376, 377, 379, 380, 384, 386-390
 holder, 358, 381
 Segregation coefficient, 360-362,

Silane oxygen deposition, 403, 404
 Silane reduction process, 417
 Silica, 17
 Silicide, 392, 400, 420, 423-425
 Silicide formation, 534
 Silicide process, 534, 536
 Silicide sintering, 424
 Silicon (Si), 44-46, 50, 55, 56, 60, 69, 73, 74, 77-81, 83, 86-87, 94, 105-108, 115, 121-124, 135, 141-143, 148, 150, 157-159, 284, 285, 292, 357, 361, 363, 364, 366-369, 371-381, 384-390, 431, 437, 444-449, 455, 460-461, 463-465, 467, 468, 473, 478, 479, 481, 482, 485-488, 490-495, 499-504
 APD, 331, 332, 334, 335
 crystal, 25
 dioxide, 16, 29, 357, 369, 392, 394-397, 402-404, 409, 417, 425, 427, 481, 482, 499
 etching, 447
 ingots, 357, 358, 360, 368
 nitride, 392, 400, 402, 405, 406, 424, 426, 449, 461, 512, 518, 534,
 reduction process, 417
 tetrachloride, 377, 378
 trench etching, 460, 461
 Silicon-on-glass, 510
 Silicon-on-insulator (SOI), 195, 210, 212, 225, 499, 500, 501
 Silicon-on-nitride, 212
 Silicon-on-oxide, 212
 Silicon-on-sapphire, 510
 Silicon-on-spinel, 212
 Silver, 15
 SIMOX, 514
 Simple cubic lattice, 19
 Single drift IMPATT, 262
 Single-electron memory cell (SEMC), 219, 220, 226
 Single-frequency lasers, 312
 Single-wafer etcher, 458, 459
 Si-O bonds, 172
 SiO₂, 437, 438, 447-450, 455, 461, 464, 465, 467, 496, 497, 499, 503, 518, 525, 534
 SiO₂-Si MOS capacitor 160, 169, 199
 Si-Si bond, 172
 Slurry, 422, 423
 Smart cut, 500, 514
 Small-scale integration (SSI), 507
 Small molecules, 296
 Smart IC, 214
 Sn, 370, 384
 Snell's law, 290
 SnO₂, 345, 348, 414
 Sodium, 19, 173
 SOI (see silicon on insulator)
 SOI devices, 212, 225
 SOI integration, 535
 SOI technology, 535, 539
 Solar cell, 3, 4, 280, 281, 284, 323, 336-339
 Solar constant, 336
 Solar radiation, 336, 338, 353
 Solar spectral irradiance, 337
 Solid solubility, 418
 Solid-phase epitaxy process, 492
 Solubility, 374
 SONOS transistor, 222
 SOS, 212
 Source, 180-184, 186-192, 194, 195-198, 200-207, 210-214, 218, 223
 Space charge, 84-86, 88-91, 95-97, 102, 115, 157
 density, 85, 233, 266
 neutrality, 88
 effect, 43, 71, 72, 77, 446
 region, 84, 86, 121, 137,
 Space-charge-limited current, 175
 Specific contact resistance, 228, 238, 239
 Spectrum splitting, 341
 Spectral width, 287
 Spiking, 417-420
 Spiral inductor, 510, 511, 539
 Spontaneous emission, 280-283, 306, 308
 Spontaneous radiation, 286, 291, 319
 Spray etching, 447
 Sputter etching, 451
 Sputtering, 414-417
 Stacking fault, 373, 374
 Staebler-Wronski degradation, 346
 Stainless steel, 493
 Static random-access memory (SRAM), 214, 215, 216, 225, 227
 Steel industry, 1, 2
 Step coverage, 402-404, 406, 409, 416, 417, 425
 Step index fiber, 300
 Step-and-repeat projection, 432, 433
 Step-and-scan projection lithography, 432
 Stimulated emission, 280-283, 303, 305, 319
 Stochastic space charge, 446, 462
 Storage capacitor, 160, 214, 227
 Storage cell, 214, 219
 Storage time delay, 142
 Strained layer, 357, 385-388
 epitaxy, 384, 385, 388
 superlattice, 386, 387
 Stripe geometry laser, 312
 Strong inversion, 163-166, 168, 193, 199
 Submillimeter wave band, 259, 277
 Substitutional sites, 390, 491
 Substrate bias, 181, 188, 191, 192, 196, 197
 Substrate doping, 163, 180, 186, 189, 192, 197
 Subthreshold characteristic, 187, 198, 199, 200, 210, 211
 Subthreshold current, 186, 187, 198, 199
 Subthreshold swing, 197, 194, 199, 200, 211
 Superlattice, 317, 318, 384, 386, 387
 APD, 331, 332, 334
 Surface concentration, 470, 471, 473, 474, 475, 477-482, 501-503
 Surface depletion layer, 163, 164
 Surface depletion region, 162-165, 168, 182, 183, 193, 193, 199
 Surface orientation, 371, 372, 388
 Surface potential, 163, 164, 166, 171, 172-173, 178, 183, 187, 191, 196

Surface recombination, 62, 65, 80, 148
 Surface recombination velocity, 62, 65, 80
 Surface states, 62, 231, 255
 Surface-emitting infrared InGaAsP LED, 301
 Switching, 82, 110, 120, 123, 124, 137, 141, 142, 149, 150, 154, 155, 159
 circuit, 111, 141
 time, 120, 124, 141
 transients, 141, 142
 Synchrotron, 445
 radiation, 445

 TaN, 414
 Ta₂O₅, 509, 534, 535, 540
 Tandem solar cell, 343-346
 TaSi₂, 424
 Taylor series, 246
 Technology drivers, 10, 11, 12
 TED, 265-269, 277, 279
 Television, 529, 536
 Tellurium, 369, 467, 488
 Temperature effect, 107, 310
 Tensile stress, 406
 TEOS, 402, 403, 404, 406
 deposited SiO₂, 406
 Ternary compound, 17, 18
 Terrestrial energy source, 336, 353
 Tetraethylorthosilicate, 402
 Tetrahedron bond, 22, 23
 Thermal energy, 43, 53, 56, 81
 Thermal equilibrium condition, 82, 83, 92
 Thermal-expansion-coefficient mismatch, 386, 387
 Thermal oxidation, 392-396, 425, 426, 467, 481, 492
 Thermal oxide, 392
 Thermal velocity, 44, 46, 53, 61, 73, 81
 Thermionic current, 273
 Thermionic emission, 175, 234, 235, 236, 238, 239, 255, 274, 277
 Thermionic emission process, 43, 68, 69, 81
 Thickness, 371, 372, 381, 383, 385, 387, 388, 391
 Thickness variation, 371, 372, 381
 Thin film transistors (TFT), 210
 Thin oxide growth, 400
 Three-dimension structure, 213, 214
 Threshold current density, 307, 309, 310, 315, 317, 319, 322
 Threshold energy, 435, 436
 Threshold field, 265, 267
 Threshold voltage, 195-198, 200, 211, 215, 216, 218, 219, 221, 223, 225-227, 246, 251, 252, 255, 256, 257, 483, 499, 500, 501, 503
 adjustment, 483, 499, 500
 control, 196
 shift, 196, 215, 218, 221, 227
 Throughput, 431, 437, 441, 443, 444, 446, 459, 462-464
 Thyristor, 2, 3, 4, 82, 123-124, 149-155, 159, 207, 366
 THz radiation detection system, 273
 Tile micrograph, 534
 Tilt angle ion implantation, 498, 501
 TiN, 196, 197, 414
 Tin, 462
 TiO₂, 414, 534, 535
 TiSi₂, 423, 427, 534
 Titanium nitride (TiN), 400, 417, 419, 425
 TiW etch chamber, 459
 TiWN, 532
 Top-gate structure, 210, 211
 Total capacitance, 166, 227
 Total conduction current density, 55
 Total internal reflection, 290, 291, 321
 Trace impurity, 374
 Transceiver, 524
 Transconductance, 139, 185, 186, 194, 200, 202, 223, 245, 247, 248, 253, 277, 278
 Transfer process, 428, 431, 437, 438
 Transferred electron devices, 258, 265, 268, 279
 Transferred-electron diode, 3, 4
 Transient behavior, 82, 108, 110, 111, 120, 122, 141
 Transient time, 110-111
 Transistor action, 3, 123, 124, 126
 Transit time, 140-141, 148-149
 Transit time domain mode, 258, 266, 269
 Transmission coefficient, 71, 81, 270, 271
 Transmutation, 366
 Transverse field, 181
 Traps, 411
 Tree of disorder, 490
 Trench, 413, 515, 521, 527, 536-540
 capacitor, 460
 isolation, 7, 9, 208, 225, 515, 520, 521, 538, 540
 type capacitor, 536
 Triac (triode ac switch), 155-156
 Trichlorosilane (SiHCl₃), 358, 377
 Trimethylgallium, 380, 381, 384
 Trimethylaluminum Al(CH₃)₃, 380, 381
 Triode reactive ion etcher, 456
 Tris (8-hydroxy-quinolino) aluminum, 297
 Tungsten, 19, 441, 461-462
 filaments, 493
 thermionic-emission cathode, 441
 -silicon Schottky diode, 234, 236
 Tunnel diode, 3, 4, 258, 260, 261, 270, 273, 277, 278
 Tunneling current, 238
 Tunneling effect, 111-115
 Tunneling phenomenon, 69
 Tunneling probability, 238
 Tunneling process, 112, 215, 219
 Turn-off time, 110, 142
 Turn-off transient, 141-142
 Turn-on transient, 141-142
 Twin tub, 520, 521, 540
 Twin well, 206, 466
 Two-step source/drain junction, 201
 Two-zone cathode contact, 267
 Types of MOSFET, 185, 186, 194

 ULSI, 507, 511, 524, 532, 537, 539
 chips, 507
 circuit, 406-408, 414, 421, 424,

425, 516
 Ultralarge scale integration
 (ULSI), 400, 437, 451
 Ultrahigh vacuum system, 388
 Ultrahigh-vacuum chamber, 381
 Ultrapure polycrystalline, 363
 Ultrashallow junction
 formation, 534
 Ultrathin oxide, 505, 534
 Ultraviolet, 280, 281, 286, 291,
 U-MOSFET, 223
 Unipolar resonant tunneling
 transistor, 258, 273, 274, 277
 Unit cell, 17-20, 41
 UV irradiation, 219
 UV light, 437

 Vacancy, 372, 373, 468, 469, 476,
 477, 480
 diffusion, 468, 469
 mechanism, 468, 477
 Vacuum level, 117, 119, 160, 169
 Vacuum system, 229
 Valence-band discontinuity, 147
 Valence bond, 15, 22
 Valence electrons, 22, 25, 29, 34
 Valley current, 260, 261, 273
 Vapor pressure, 368, 380, 388
 Vapor-phase epitaxy, 377

 Vapor-phase HF, 449
 Varactor, 98-99
 Variable reactor, 98
 Variable-shaped beam, 442
 Vector scan, 442-443
 Velocity saturation region, 253
 Vertical-cavity surface-emitting
 laser (VCSEL), 316, 317
 Vertical isolation, 511, 512, 517
 Via plug, 421
 Video recording, 302, 319
 Viscosity, 437
 Visible LED, 286, 291-293, 319
 VLSI, 507, 511
 V-MOSFET, 223
 Volatile memories, 528
 Volume charge density, 173, 174,
 193, 199
 Volume defects, 372, 374, 388
 V-shaped groove, 223

 Wafer, 505, 506, 507, 512, 514, 518,
 524, 525, 532, 536, 539,
 Wafer shaping, 357, 370, 371, 388
 Water molecules, 396
 Water vapor, 393-398
 Wavelength, 431-434, 437, 439, 440,
 444, 459, 460
 Weak inversion, 163

 Wet chemical etching, 428, 447, 451,
 460, 463, 464
 Wet photoresist stripping, 438
 WF₆, 417
 White LED, 291, 298, 319
 WN, 414
 Work function, 68, 81, 117, 160,
 161, 165, 169, 170, 172, 173, 181,
 188, 189, 192, 193, 196, 197,
 229, 230, 235, 256
 Work function difference, 169, 170,
 172, 196, 197
 WSi₂, 424

 Yellow light, 437
 Yield, 434, 435, 436, 440, 459, 464
 Yield strength, 374, 389

 Zinc (Zn), 18, 28, 29, 369, 380, 381
 Zinc diffusion, 479
 Zincblende lattice, 20, 21, 22, 27, 28
 Zinc sulfide, 330
 ZnAs₂, 467
 Zn-Ga-As alloys, 467
 Zn (see Zinc)
 ZnO, 414
 ZnSe, 293
 Zone-refining technique, 365